CRANFIELD UNIVERSITY

GREG FLITTON

EXTENDING COMPUTER VISION TECHNIQUES TO RECOGNITION PROBLEMS IN 3D VOLUMETRIC BAGGAGE IMAGERY

SCHOOL OF ENGINEERING

PhD THESIS

Academic Year 2011-2012

Supervisor: Dr. Toby Breckon

January 2012

CRANFIELD UNIVERSITY

SCHOOL OF ENGINEERING

PhD THESIS

Academic Year 2011-2012

GREG FLITTON

Extending Computer Vision Techniques To Recognition Problems In 3D Volumetric Baggage Imagery

Supervisor: Dr. Toby Breckon

January 2012

This thesis is submitted in partial fulfilment of the requirements for the Degree of Doctor of Philosophy.

Abstract

We investigate the application of computer vision techniques to rigid object recognition in Computed Tomography (CT) security scans of baggage items. This imagery is of poor resolution and is complex in nature: items of interest can be imaged in any orientation and copious amounts of clutter, noise and artefacts are prevalent.

We begin with a novel 3D extension to the seminal SIFT keypoint descriptor that is evaluated through specific instance recognition in the volumetric data. We subsequently compare the performance of the SIFT descriptor against a selection of alternative descriptor methodologies. We demonstrate that the 3D SIFT descriptor is notably outperformed by simpler descriptors which appear to be more suited for use in noise and artefact-prone CT imagery.

Rigid object class recognition in 3D volumetric baggage data has received little attention in prior work. We evaluate contrasting techniques between a traditional approach derived from interest point descriptors and a novel technique based on modelling of the primary components of the primate visual cortex.

We initially demonstrate class recognition through the implementation of a codebook approach. A variety of aspects relating to codebook generation are investigated (codebook size, assignment method) using a range of feature descriptors. Recognition of a number of object classes is performed and results from this show that the choice of descriptor is a critical aspect.

Finally, we present a unique extension to the established standard model of the visual cortex: a volumetric implementation. The visual cortex model comprises a hierarchical structure of alternating simple and complex operations that has demonstrated excellent class recognition results using 2D imagery. We derive 3D extensions to each layer in the hierarchy resulting in class recognition results that significantly outperform those achieved using the earlier traditional codebook approach.

Overall we present several novel solutions to object recognition within 3D CT security images that are supported by strong statistical results.

For fun...

Acknowledgments

I would like to thank my supervisor, Toby Breckon, for his efforts and guidance during my research. I simply don't know where he gets his energy from but if he could bottle it...

I would also like to thank Ousha for her support and little Amelie, whose early arrival on the scene disrupted things at the end, though she did provide me with some company during the nights prior to submission. Ma and Pa gave me all the help I needed over the years...

Thanks are due to Najla who always entertained my ridiculous thoughts, Ioannis, who endured my funny pyramid phase over tea and sticky buns, and Mike, who provided me with numerous cups of tea on my way home.

I also wish to thank Reveal Imaging Technologies for their assistance over the years.

Neil needs a mention, not least because he provided me with my first publication, courtesy of MBR and RAF Search and Rescue.

Oh, and not forgetting...

This project is funded under the Innovative Research Call in Explosives and Weapons Detection (2007), a cross-government programme sponsored by a number of government departments and agencies under the CONTEST strategy.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This work considers the problem of object recognition within complex 3D Computed Tomography (CT) imagery arising from security scans of baggage items. Conventional security scans use 2D X-ray technology to aid an operator in the detection of items of interest but the recent introduction of CT scanners has provided a source of 3D imagery for interpretation - a form of imagery that is common within the medical imaging domain. We take this 3D imagery and examine computer vision techniques for automatic recognition of objects of interest with the intent of improving transport security.

1.1 Automatic recognition of items in baggage

The primary objective of this research is to implement and evaluate novel computer vision techniques for the automatic recognition of rigid items within CT-baggage imagery - a challenge that has not been attempted in prior work. We must ensure a high true-positive rate of detection so that items of interest are not missed whilst at the same time maintaining a low false-positive rate so that the impact of an overly cautious detection algorithm is minimal. If we can demonstrate automatic recognition of objects using CT scanners with a high true-positive rate and low false-positive rate then this would enhance security of transport infrastructure throughout the world.

Our research question is "can computer vision techniques be applied to 3D CT imagery of baggage items in order to recognize potential items of interest with a high true-positive rate and low false-positive rate?".

1.2 CT imagery and computer vision techniques

Conventional security scanning uses 2D X-ray technology. An alternative imaging technique has recently been deployed in the security-scanning environment: Computed Tomography (CT). This technology has been successfully used for medical imaging for many years where a 3D image is created from a series of cross-sectional slices allowing improved visualization of medical conditions. Figure 1.1 shows an example of this technology in the baggage-imaging domain where we can see individual slices through a cluttered bag (Figure 1.1a) being combined and viewed as a 3D volume (Figure 1.1b).

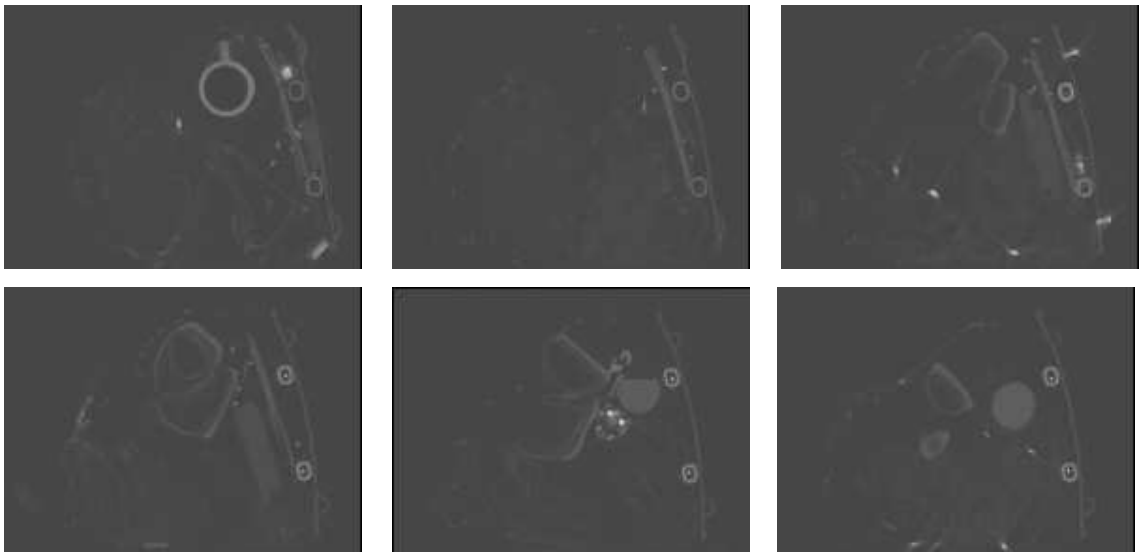
The use of CT imaging in the security environment has arisen through dual energy variants which excel at detection of explosive materials but also present imagery which does not suffer from self occlusion, objects can be segmented from the 3D image in their entirety, potentially allowing a detailed analysis of the entire bag to be made. Our research makes use of such imagery and develops explicit 3D extensions of current state of the art approaches within the field of computer vision to tackle the challenge of automated object recognition in cluttered CT-baggage imagery, typical of that found in an aviation-transport situation.

Explicitly we explore

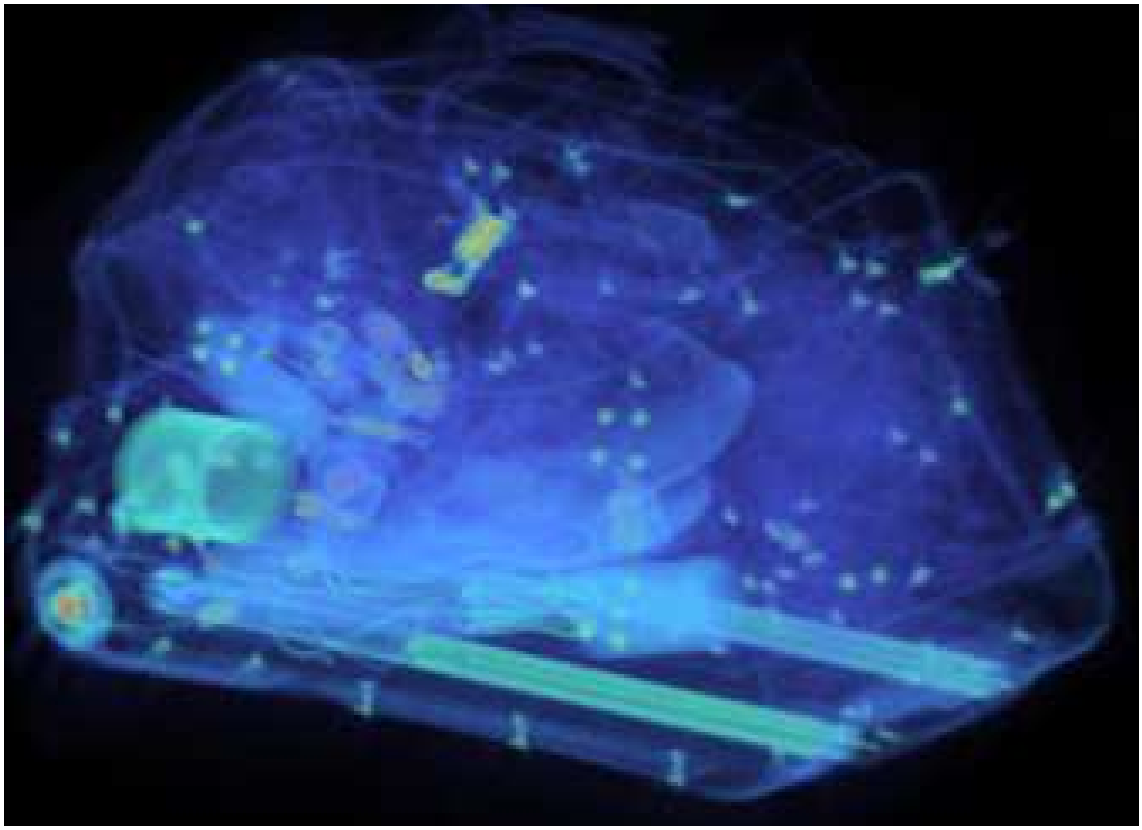
- A 3D extension of rigid object detection using a 3D extension of SIFT (Lowe, 2004).
- A comparison of a range of 3D volumetric point descriptors for both individual object detection and bag-of-features driven object classification (Sivic and Zisserman, 2003; Csurka et al., 2004).
- The 3D extension of an established visual cortex model approach, driven by Gabor filter response features, for the same 3D classification task (Mutch and Lowe, 2008).

1.3 Improving transport security

Detailed examination of baggage items in transport security infrastructure has been a requirement for many years for a variety of reasons including terrorist threats and smuggling of contraband items. In the case of airport security there are three distinct sources of threat. Large items of baggage that are transported in the aircraft hold can be used to contain explosive devices that can be triggered using a simple timer (Lockerbie disaster, 1988; Air India Flight 182, 1985). Small luggage that can be



(a) Example 2D CT slice images



(b) 3D volumetric image can be viewed from any point

Figure 1.1: CT imagery of baggage item: from slices to 3D volume

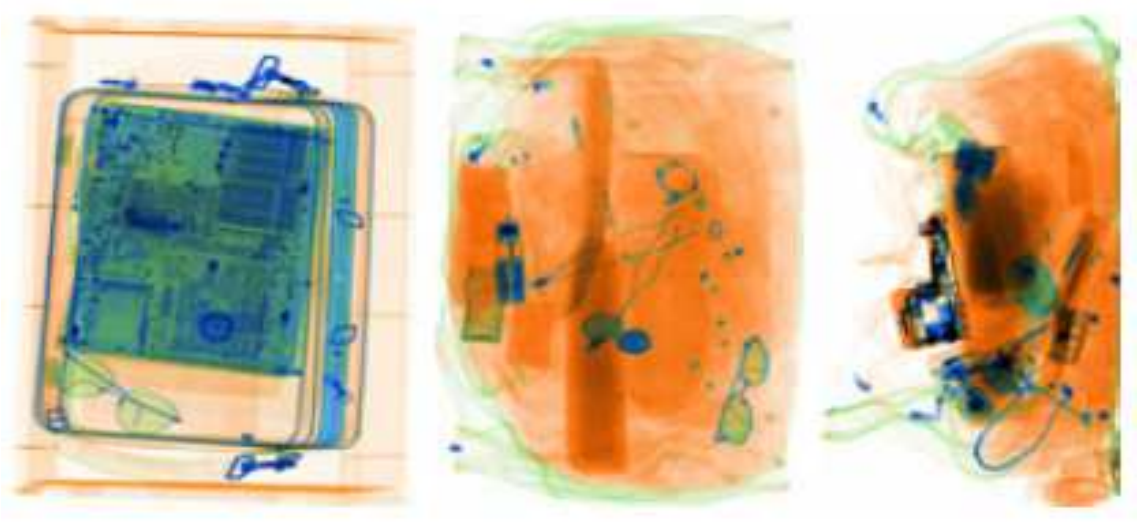


Figure 1.2: Using false colour to enhance 2D X-ray image (taken from Baştan et al., 2011)

carried on to the aircraft can be used to contain explosives, weapons or contraband items. Similarly, items can be secreted about the passenger in various ways in order to smuggle items through the security checks.

Detection of explosives can be achieved using a number of techniques relating to the physical characteristics of the chemicals employed. Similar approaches can be used for drugs but in both these cases there is no specific shape that can be recognized. In some cases drugs can be recognized from an X-ray by alterations that have been made to known items (voids are created inside objects into which the contraband is concealed) but in general this is a complex task.

Identifying items within baggage is achieved using 2D X-ray scanners that require a human operator. Automatic detection of explosives is possible through examination of the response of the materials within the baggage to the X-ray energy but recognition of other items relies on the skill and experience of the operator. The operator is assisted in this process through the application of false colour to the 2D scalar X-ray imagery. Figure 1.2 shows an example where we can see the false colour applied to the imagery in order to identify organic (food, paper, fabric: orange), inorganic (plastics, metals: blue) and mixed materials (green).

Conventional X-ray techniques suffer from the problem of self occlusion - the 2D X-ray projection cannot separate individual items as they overlap. This inhibits the identification of explosive material but also makes the recognition of objects more complex. The human operator has a number of difficult visual tasks - recognition of known weapons (guns, knives); recognition of potential weapons (razor blades); identification of contraband (drugs, alcohol); recognition of improvised explosive

devices (IEDs). These tasks have to be achieved within a limited time frame (6-10 seconds is typical) though if something suspicious is found more time can be allocated to make a final decision.

At present there is little use made of computer vision techniques in deployed baggage scanners to automatically detect items of interest to the security agencies. Of the research that has been published on recognition using 2D X-ray imagery, the self-occlusion problem is often cited as a key area that inhibits detection performance.

In this research we have an opportunity to examine automatic detection of items in baggage items through the use of a different imaging paradigm - Computed Tomography, as illustrated in Figure 1.1.

1.4 Contribution to knowledge

Our ultimate goal is accurate object class recognition with a low false positive/false-negative rate for a class of imagery (e.g. cluttered 3D CT imagery) which has received very little attention within the object recognition domain in prior work (Bi et al., 2009). To achieve this, we move from specific instance recognition to full class recognition in a series of steps aimed at building a firm basis for subsequent analysis. Recognition is performed on a variety of objects throughout the thesis with handguns (weapon) and bottles (cf. liquid explosive container) being used in determining class recognition performance.

Through these steps the work presented in this thesis extends the current state of the art within the automatic recognition domain, with the following notable contributions.

- We demonstrate specific-instance object recognition in 3D CT imagery through the use of a 3D extension to the SIFT descriptor (Lowe, 2004) extending prior 3D SIFT work looking only at the problems of image registration/3D panorama creation (Allaire et al., 2008; Ni et al., 2009).

- We compare the performance obtained using our novel 3D SIFT descriptor with other descriptors, including an extension of the established RIFT descriptor (Lazebnik et al., 2005) to 3D, that are significantly simpler in concept, and thus computationally more efficient in implementation. We show that such descriptors can produce better recognition than the 3D extension of the seminal SIFT work (Lowe, 2004). Furthermore we also note that, within the context of 3D CT baggage imagery data, feature matching using the distinction method of Lowe (2004) is outperformed by taking a fixed percentile of

the matches that have been ordered by Euclidean distance. This is in contrast to the established and accepted 2D methodology of Lowe (2004). This extends prior work in 2D descriptor comparison (Mikolajczyk and Schmid, 2005) and specifically extends the work of (Lowe, 2004; Lazebnik et al., 2005).

- We examine object class recognition through development of a bag-of-features approach (Sivic and Zisserman, 2003; Csurka et al., 2004) using a number of codebook assignment methods based on the set of evaluated descriptors. Again this shows 3D SIFT is outperformed by other descriptors, notably local density and density-gradient histograms, supporting the conclusions of our earlier comparative study. Furthermore the results of these tests show a relative under-performance in the detection of feature sparse objects (i.e. those with little density variation, for example bottles containing liquids) when compared to feature rich (i.e. complex objects, e.g. handguns) indicating that the key-point methodology employed is dependent on the generalized feature density of the objects considered.

- Finally we develop a full 3D extension to the visual cortex standard model (Mutch and Lowe, 2008) that has shown considerable promise in 2D recognition and notably requires a minimal set of training data as is typified by the problems we deal with within this application space. Successful detection of both handguns and bottles is high (outperforming our earlier bag-of-features study). This indicates that this approach may be a more generalized solution to object class recognition in CT baggage imagery given both its detection performance and limited training data requirements.

1.5 Outline of the thesis

In Chapter 2 we review the literature relevant to our area of research. We cover various 2D recognition techniques and their extensions into 3D both for specific instance and class recognition. We review published work in the area of baggage scanning and then examine recognition in the 3D medical imaging community as this area is the main source of 3D related literature. In Chapter 3 we discuss the datasets used for our analysis. In particular the basic image processing, re-scaling and re-sampling are outlined and examples of the types of items captured are given. In Chapter 4 we examine specific instance recognition through a 3D extension of the Scale Invariant Feature Transform (SIFT). We locate reference objects in unseen baggage and report detection rates. We then extend this work

(Chapter 5) by examining the performance of the SIFT descriptor against other descriptors following concerns with the SIFT rotation invariance methodology raised in matching performance. A larger set of objects is used and an investigation made to determine a reliable form of feature matching.

We then move on to class recognition in Chapter 6. We explore recognition through the 'Bag of Features' codebook approach. We use a variety of interest point descriptors and investigate methods of assignment to the codebook to determine the best recognition solution.

A new recognition paradigm is then investigated through direct modelling of the visual cortex (Chapter 7). Existing work in this area focuses on 2D implementation but we fully extend the methodology into 3D and examine class recognition performance on handguns and bottles.

Finally we summarize and discuss the results of the research and identify areas for future research work within this domain (Chapter 8).

1.5.1 Prior peer reviewed publication

To date, work presented in this thesis has been presented in the following peer reviewed publications / submissions:

- Object Recognition using 3D SIFT in Complex CT Volumes (G. Flitton, T.P. Breckon, N. Megherbi), *In Proc. British Machine Vision Conference*, pp. 11.1-12, 2010.

- A Comparison of 3D Interest Point Descriptors with Application to Airport Baggage Object Detection in Complex CT Imagery (G. Flitton, T.P. Breckon, N. Megherbi), *Pattern Recognition* (under review)

- A 3D Extension to Cortex Like Mechanisms for 3D Object Class Recognition (G. Flitton, T.P. Breckon, N. Megherbi), *2012 IEEE Conference on Computer Vision and Pattern Recognition*

Chapter 2

Literature review

At the outset of this review it is worth noting that there is very little published work in the area of threat detection in 3D CT-baggage imagery. This review of the available literature will record the limited existing work in the field of automatic threat detection in baggage imagery for both CT and 2D X-ray data but will examine other areas of research in order to establish a sound base for our work.

Medical research is the prime driver for 3D voxel-based imagery and we will discuss the computer vision techniques that have arisen from investigations in this area, in particular a 3D extension to the seminal SIFT algorithm (Lowe, 2004). It seems sensible to consider progress in automatic 3D medical image analysis prior to investigation of the baggage imagery.

We will also report on a different recognition paradigm - direct modelling of mammalian vision systems. This approach has yielded some excellent results in 2D (Mutch and Lowe, 2008; Serre et al., 2005b) and its application to 3D imagery for the task of object class recognition will be explored.

We first review the techniques and methodologies used for both known object and object class recognition in 2D and 3D. Many techniques that are used in 3D medical image analysis start life as 2D variants and so we will begin the discussion by highlighting the 2D techniques and methodologies that may be useful for both specific instance and class recognition. For further insight and background within the generalized form of object recognition the reader is referred to Szeliski (2010) and Felzenszwalb et al. (2010).

2.1 Specific instance recognition

Recognition of a known object in a previously unseen image is something humans take for granted. We begin by examining several computer vision techniques that

have been created to address this problem.

2.1.1 Overview

The detection of a known item in a given scene is one of the tasks that the human vision system performs extremely well. The ability to perform this task on images containing a specific item in a complex scene has been investigated in the computer vision community since its inception (Ballard and Brown, 1982) and has more recently become of increased interest - interpretation of images taken from smart-phones to aid shopping tasks, for instance.

One technique is the formulation of a mathematical model of an item of interest followed by matching, through geometric alignment, between model and image containing the actual item. Lowe (1987) introduced the SCERPO vision system that performed recognition on gray-scale imagery using 3D wire frame models. Line segments are extracted from the target image. A subset of these are used to hypothesize an orientation for the 3D reference model such that, when projected into 2D, a consensus for the proposed match is achieved. Figure 2.1 shows an example from Lowe (1987) showing matches between a wire frame model of a razor and an image containing razors in various orientations with partial occlusion prevalent. These approaches work well for controlled situations (i.e. controlled lighting conditions with no background clutter) but do require the creation of an accurate model of the reference item. Relying on line segments restricts this recognition technique to mainly man-made objects. An improvement can be made by using other salient aspects of the reference object for recognition.

One such approach is the use of interest feature points to locate a reference object in an image.

Lamdan et al. (1988) began using interest points for the detection of 3D objects in 2D binary images. Points of interest were identified from "sharp convexities and deep concavities" in the outline, as shown in Figure 2.2 (taken from Lamdan et al., 1988). The recognition algorithm takes two steps. Firstly, interest points are matched in groups of three, so that an estimate of the transformation required to back-project the model onto the image can be made. The reference item is then transformed and a verification step is made by comparing the edges of the transformed reference item and the edges in the scene. If this verification step fails then a different set of matches is used in an attempt to find a plausible result. This approach generally follows an interpretation tree methodology (Grimson and Lozano-Perez, 1987; Fisher, 1989, 1994) - plausible matches are connected in a tree structure to swiftly establish a 'best match' that can then be accepted or rejected.

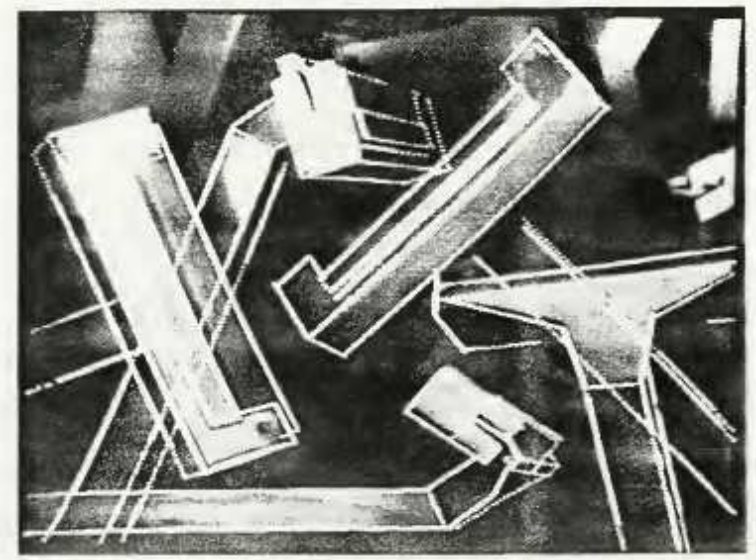


Figure 2.1: Lowe matching

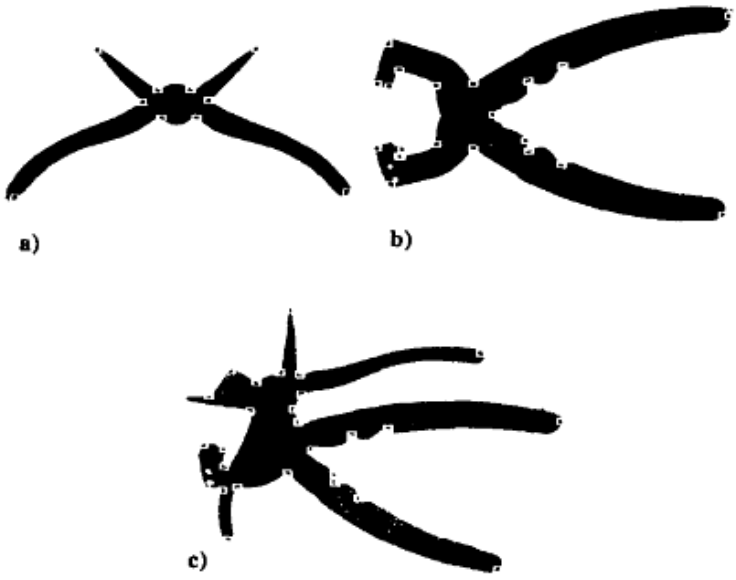


Figure 2.2: Interest points from Lamdan et al. (1988)

For many 2D applications, perspective distortion will appear to warp objects within the imagery. In many approaches this effect is modelled as an affine transform which allows for translation, rotation, scale and shear in the image. In our 3D imagery we do not need to account for perspective changes so the requirement for interest point detectors that are tuned to affine invariance are not required.

As points of interest have proven to be useful in the recognition of objects, the methodology that derives their location has received much attention. The "sharp convexities and deep concavities" of Lamdan et al. (1988) needs a more formal approach. We now discuss various methodologies for locating and describing interest points in images.

2.1.2 Interest point detection and location

Many types of interest point detector have been created for operation on 2D imagery with numerous applications (stereo correspondence, image stitching, recognition, tracking, registration, simultaneous localization and mapping). It is often a requirement to recognize objects regardless of the perspective distortion that will vary their appearance. Size, for instance, is often taken into account using a scale-space pyramid (Lowe, 2004) where each image layer is processed as a separate entity in the recognition methodology.

One of the most successful approaches is the corner detector of Harris and Stephens (1988) which locates points of interest by examining the auto-correlation function for a given patch of pixels. This detector is invariant to rotation though not to scale. It has been used in numerous applications (Schmid and Mohr, 1997; Tommasini et al., 1998) and has been extended into 3D for the purpose of identifying points of interest in spatio-temporal imagery (Laptev, 2005). A close relation of this detector was developed by Shi and Tomasi (1994) for the primary aim of feature tracking in video.

The FAST detector of Rosten and Drummond (2006) compares a central pixel to the surrounding pixels. The central point is declared an interest point if the surrounding pixels covering an arc of 270° are all higher or all lower by a threshold value than the central point.

Another approach uses the Hessian matrix which describes local curvatures to a point.

$$\text{Hessian Matrix: } H(I(x,y)) = \begin{bmatrix} \frac{\partial^2 I}{\partial x^2} & \frac{\partial^2 I}{\partial x \partial y} \\ \frac{\partial^2 I}{\partial x \partial y} & \frac{\partial^2 I}{\partial y^2} \end{bmatrix} \tag{2.1}$$

If the determinant of the Hessian is calculated for all points in the image then

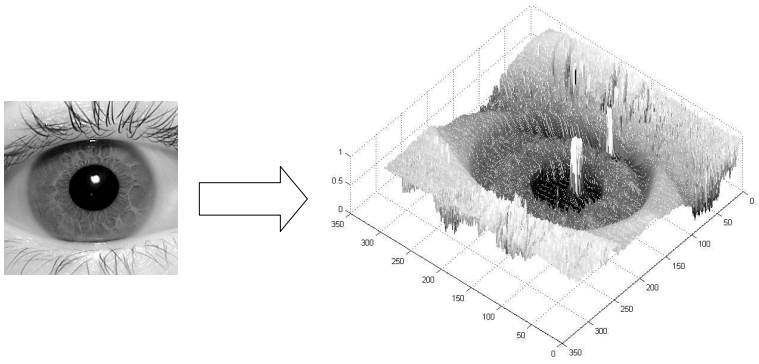


Figure 2.3: Consideration of image from topographical viewpoint

local maxima are indicative of areas of interest.

An alternative to locating points associated with corners is to find blobs or regions in an image that are brighter or darker than the surrounding neighbourhood. The Difference of Gaussian (DoG) approach used in the SIFT methodology (Lowe, 2004) is a form of blob detector and is discussed in detail in Section 2.1.5.

Another blob detector was presented by Matas et al. (2004) as Maximally Stable Extremal Regions (MSER). Closely related to watersheds (Vincent and Soille, 1991), the 2D image is considered topographically with pixel values forming peaks and troughs (see Figure 2.3). The MSER algorithm finds regions in the image by forming connected components as a threshold is varied from minimum to maximum pixel value. The area of each connected component is recorded as the threshold changes. This variation is analyzed and regions are declared maximally stable for threshold values that result in a minima of the rate of change of the connected component area. Extensions of MSER into 3D have been reported (Donoser and Bischof, 2006; Riemenschneider et al., 2009) for the purpose of action recognition in spatio-temporal 3D imagery.

An evaluation of interest point detectors in 3D has recently been published (Yu et al., 2011). Although mainly analyzed on synthetic data the results indicate that the 3D version of MSER (Donoser and Bischof, 2006; Riemenschneider et al., 2009) produces interest points that are more easily matched in noisy imagery, echoing prior 2D work (Mikolajczyk et al., 2005).

2.1.3 Interest point description

Following the location of an interest point, a method is required that describes it in a manner that allows accurate matching between images. An important aspect of objects being described is that they are textured - a useful requirement are points of interest located within the object boundary rather than on it. Points of interest that

Figure 2.4: Use of SIFT descriptor for object recognition (taken from Lowe, 1999)

are within an object are less likely to be altered by adjacent objects in a cluttered scene and hence it is more probable that the point will be accurately described from one scan to another.

One of the most popular descriptors is the Scale Invariant Feature Transform (SIFT). The development of the seminal SIFT descriptor began with Lowe (1999) where the location of objects in a scene was demonstrated, as shown in Figure 2.4. Refinement to the methodology (Lowe, 2004) introduced smaller scale space steps in the scale-space pyramid, sub-pixel estimation in the location of points of interest and multiple descriptors for each location if there was not a clear dominant direction in the gradients at the keypoint. The SIFT methodology will be described in detail in section 2.1.5.

Various descriptors including SIFT are discussed in Mikolajczyk and Schmid (2005) with the conclusion that the SIFT-based descriptors have the best performance under various image transformations (rotation; scale; blur; viewpoint; illumination change).

Not included in the review of Mikolajczyk and Schmid (2005) is the SURF descriptor, introduced by Bay et al. (2008) as a scale and rotation invariant descriptor, claimed to achieve higher performance than SIFT (having a superior recall for a given precision when finding a known point in a new image) and be faster to compute. Points of interest are located as local maxima in the determinant of an approximate

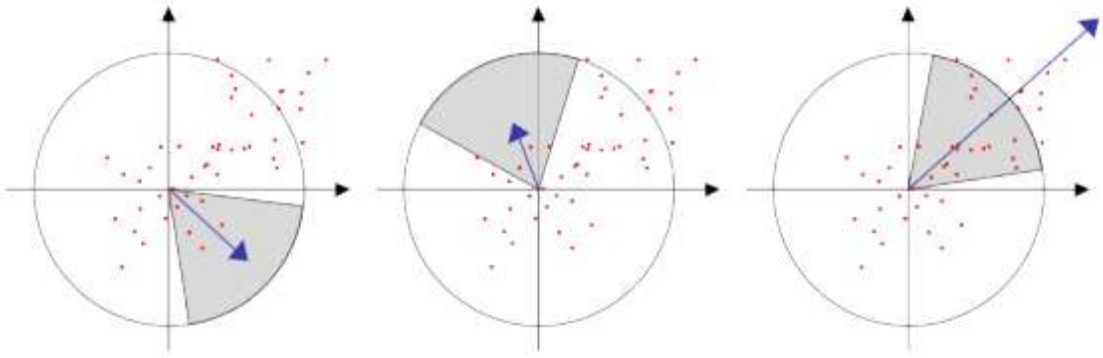


Figure 2.5: Dominant orientation generation in SURF (taken from Evans, 2009)

Hessian matrix dubbed the "Fast Hessian Detector". Orientation invariance at each keypoint is achieved by first calculating Haar-wavelet responses in each direction (x, y) for each pixel in the neighbouring region. A sector window of angle $\pi/3$ is then applied at the keypoint and rotated. The wavelet responses within the sector window are summed to create an orientation vector. The largest vector generated as the sector window alters its position is taken as the dominant orientation for the keypoint, as shown in Figure 2.5. The dominant orientation is used to re-orientate the keypoint region and then a description window is created around the centre comprising a 4×4 region grid comprising 5×5 pixel cells. The Haar wavelet responses within each region, (d_x, d_y), are summed to record four distinct values: $\sum d_x$, $\sum d_y$, $\sum \mid d_x \mid$, $\sum \mid d_y \mid$. The result is a descriptor comprising 64 elements (4×4 grid, 4 elements per grid) that is subsequently normalized to unity.

2.1.4 Interest point matching / object location

Following interest point location and description, the recognition task is faced with the problem of locating an object from one or more reference images, or a 3D model, in a target image. A number of approaches can be taken for this step.

Lowe (2004) used the generalized Hough transform (Ballard, 1981) to vote on the location of the object in the target image. Each SIFT descriptor maintains a local orientation relative to the centre of the reference object. If the same descriptor is located in the target image then the location of the object centre can be estimated. If at least three interest point matches agree on the same location for the object then a verification stage is entered. Using the interest point locations an affine transform is calculated that transforms the interest points from the reference object 2D image onto the target image space. A comparison is then made between the reference item and proposed target location interest points before a final recognition decision is made.

A number of methods can be used to determine interest point matches. Matching interest points with the closest Euclidean distance is one approach, though Lowe (2004) found this to be unreliable and instead looked for matches that appeared distinct (discussed in Section 2.1.5.1). Alternatives to the Euclidean distance as the match metric include the Earth Mover's Distance (Rubner et al., 2000) and Mahalanobis Distance (Mahalanobis, 1936).

An alternative to the Hough transform is to use the RANdom SAmpling and Consensus algorithm (Fischler and Bolles, 1981), more commonly referred to as RANSAC. This approach initially chooses a subset of matches at random before using them to form a transform between model and target image. A consensus is then sought between the remaining reference item keypoints and those in the target image. If such a consensus is reached then recognition is declared. A refinement to the transform can be made using the extended set of matches followed by patch-based verification. If no consensus is reached then the process is repeated. An upper limit to the number of attempts is required to account for the possible situation where the reference object is not present. The algorithm has been proven to cope well in the presence of numerous outliers and has triggered the generation of numerous enhancements (Chum et al., 2003; Chum and Matas, 2005; Sattler et al., 2009; Torr and Zisserman, 2000; Nistér, 2005; McIlroy et al., 2010).

2.1.5 Scale invariant feature transform

The 2D SIFT descriptor has been used in many applications: image stitching (Brown and Lowe, 2007), registration (Yi et al., 2008; Bustard and Nixon, 2008), recognition (Luo et al., 2007; Sivic et al., 2005; Bicego et al., 2006; Collet et al., 2009; Belcher and Du, 2009), segmentation (Feng et al., 2009), robot localization and mapping (Sim et al., 2005; Se et al., 2002; Tamimi et al., 2006). Given its importance in our work we will now give details of its operation.

2.1.5.1 2D implementation

Originally created by Lowe (1999) it was refined (Brown and Lowe, 2002) before the definitive published work was presented (Lowe, 2004). There are four main stages to the algorithm and these are shown in Figure 2.6.

The first step is to search through location and scale space for candidate interest points. This takes the form of applying a Difference of Gaussian (DoG) filtering process and retaining locations (in space and scale) of the extrema. Figure 2.7a illustrates the DoG process (taken directly from Lowe, 2004). Figure 2.7b shows

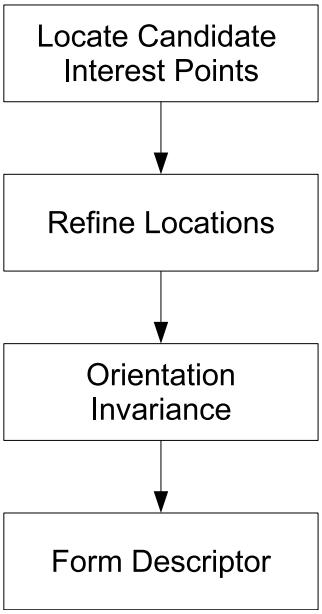
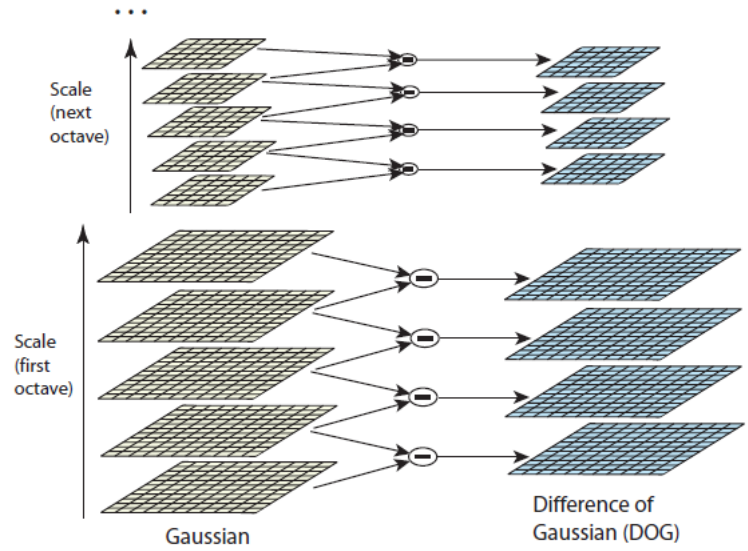


Figure 2.6: SIFT algorithm: stages to description

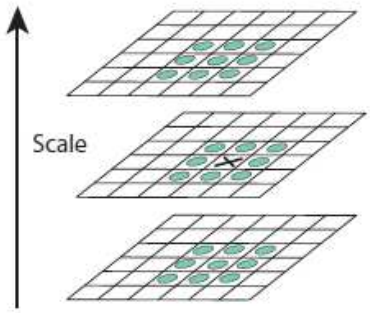
extrema are located through a comparison in scale space (pixel location and adjacent scales).

The second step starts by refining the keypoint locations to sub-pixel accuracy. Introduced by Brown and Lowe (2002) this step fits a 3D quadratic equation (for 2D position plus scale) to the candidate point then solves it to produce a refined estimate of the true location. This is followed by analyzing the refined candidate locations and removing those that are unlikely to provide stable descriptors. In particular, the method removes candidates that are located in regions of low contrast (low pixel values) or are not located at points of high curvature in all dimensions. This removes points that will be easily corrupted by noise or located on edges/ridges that will be unreliable for matching purposes. Figure 2.8 illustrates this where we can see that interest points located on edges will produce similar descriptors that will hinder subsequent matching whereas other interest points will create unique descriptors that will be easier to match.

Having refined the candidate locations the method moves on to resolving the problem of rotation. It is important that the resultant descriptors are invariant to rotation as this will allow robust matching from one image to another even if a rotation has occurred. Using the gradients from the region around the keypoint a histogram of gradient orientations is created by accumulating the gradient magnitudes into the histogram bins according to the gradient orientation. Each bin in the histogram covers a 10° sector. The histogram contributions are Gaussian weighted such that gradients closer to the keypoint have a stronger contribution. Peaks in



(a) Difference of Gaussians applied in space and scale (taken from Lowe, 2004)



(b) Locating extrema in space and scale (taken from Lowe, 2004). 'X' is an extrema relative to all the points surrounding it in the scale space.

Figure 2.7: Location of candidate extrema locations

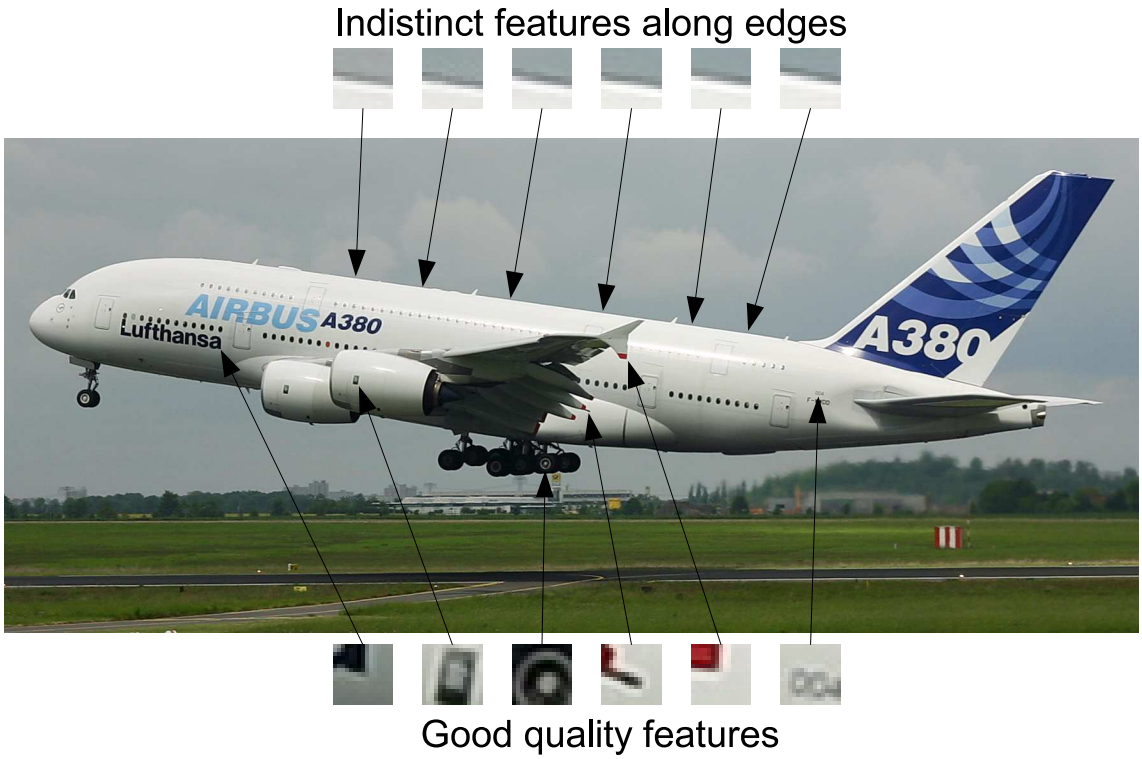


Figure 2.8: Rejecting interest points on ridges/edges

the orientation histogram indicate dominant directions of the contributing gradients. All peaks within 80% of the largest peak are defined as dominant and are used to generate a keypoint descriptor.

The final stage of the process is the generation of the keypoint descriptor. Given the calculated keypoint location, scale and orientation, the local neighbourhood at the appropriate scale is rotated and translated such that the keypoint is at an integer location. The neighbourhood gradients are then Gaussian weighted so that gradients closer to the keypoint contribute more to the descriptor. The neighbourhood region is divided into a grid structure and an orientation histogram is created for each grid region, as shown in Figure 2.9. The orientation histograms at this stage cover 45° sectors. Figure 2.9 shows a descriptor covering a 2×2 grid, each part of the grid containing a histogram of 8 elements - 32 elements in total. In the original work (Lowe, 2004) best results were achieved using a 4×4 grid leading to a descriptor of 128 elements (The blue circle indicates the Gaussian weighting applied to the gradients prior to accumulation in the orientation histograms).

When matching keypoints from one image to another, Lowe (2004) uses the ratio of match distance from the closest to the second closest neighbour - filtering to retain 'distinct' matches. This is discussed further in Section 5.3 in the context

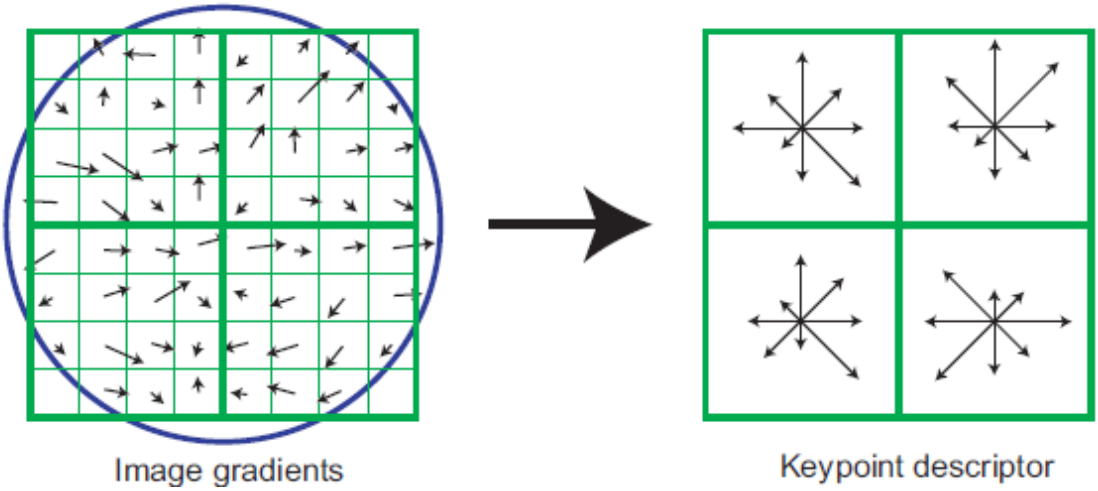


Figure 2.9: SIFT descriptor generation: grids of gradients

of the recognition problem under consideration.

2.1.5.2 Extensions to 3D

Various extensions of the SIFT algorithm into 3D have been recently presented in the literature by a number of authors (Urschler et al., 2006; Scovanner et al., 2007; Cheung and Hamarneh, 2007; Ni et al., 2009; Allaire et al., 2008). Urschler et al. (2006) used a simple extension to the descriptor for the task of registration in CT medical imagery. Keypoint locations were derived using Foerstner corners (Forstner, 1986) rather than a 3D extension to the Difference of Gaussians approach. No attempt was made at identifying rotational invariance in the descriptor as it was felt unlikely to add to the performance given that the person being scanned was not moving. The goal of the research was to use the descriptors to perform faster registration with better accuracy than an existing method - both of these requirements were achieved.

Scovanner et al. (2007) derived a form of 3D SIFT to identify types of human action in video (waving, jumping, bending etc.) and were the first to implement a form of rotation invariance. Video volumes were formed to which the 3D SIFT descriptor was applied at random points in space and time (the Difference of Gaussian approach for interest point location was not used). Clustering of the descriptors was used to form a dictionary for a bag-of-words algorithm. A support vector machine (Cortes and Vapnik, 1995) was then trained and recognition performance evaluated. Recognition of 10 types of action was achieved with an average precision in excess of 80.0%, demonstrating an improvement over other techniques.

Cheung and Hamarneh (2007) created a N-Dimensional SIFT variant to aid medical image alignment for MRI and CT scans. The DoG methodology was used to derive points of interest, though location refinement was not implemented. The methodology attempted to generalize the extension of SIFT above 2D to an N-dimensional space through the use of hyper-spherical coordinates (Schweizer, 2001) in the description of gradient directions. The resultant descriptors were used for registration of 3D MRI medical images and 4D (3D + time) CT images and performed well, provided there was little rotation between the images being considered. It was noted that as little as 10° rotation would prevent matching - we believe this is due to the use of the hyper-spherical coordinate system and will be discussed later in this section.

Ni et al. (2009) also implemented SIFT in a 3D formulation, extending the original work of Scovanner et al. (2007), for use in 3D ultrasound panoramic imagery. The DoG methodology is used to derive candidate interest points some of which are rejected if the location is on an edge or in a region of poor contrast. The SIFT descriptors are used to match between ultrasound volumetric imagery to enable the creation of a larger volume through accurate stitching. It is noted in the published work that matching performance decreases significantly as volumes are rotated relative to each other. The reason for this may lie in an error in the definition of orientation in a 3D space. This will now be discussed.

All of these approaches (i.e. Scovanner et al., 2007; Cheung and Hamarneh, 2007; Ni et al., 2009) suffer from a fundamental limitation in their consideration of orientation - the definition of orientation in 3D is incorrectly taken as the direction formed by two angles (azimuth, elevation). To correctly orientate an object in 3D requires three angles - azimuth, elevation and tilt. This error of (Scovanner et al., 2007; Cheung and Hamarneh, 2007; Ni et al., 2009) was noted by Allaire et al. (2008) and corrected to provide full orientation invariance in the descriptor. The work of Allaire et al. (2008) covered registration of medical imagery using Magnetic Resonance, Computed Tomography and Cone Beam Computed Tomography scans. The differing imagery that results required tuning of the SIFT parameters in order to maximize performance.

Riemenschneider et al. (2009) were interested in action recognition in volumetric data created from video frames. A comparison was made between the use of the 3D SIFT descriptor derived by Scovanner et al. (2007) and a descriptor based on shape context (Grundmann et al., 2008). They found that the 3D SIFT descriptor did not perform as well as the shape context method in their recognition methodology.

Niemeijer et al. (2009) used the less accurate descriptor of Cheung and Hamarneh

(2007) for registration of volumetric imagery obtained from Optical Coherence Tomography scans on the retina. Candidate features were examined using the distinction criteria of Lowe (2004) (see Section 2.1.5.1) using 0.9 as the distinction threshold.

Dalvi et al. (2010) investigated volumetric ultrasound alignment using Harris points (Harris and Stephens, 1988) on 2D image slices. They compared the performance against that obtained using the 3D SIFT approach of Ni et al. (2009) and found that the 3D SIFT approach failed to perform the registration correctly. They speculated that the reason may be that the 3D SIFT algorithm rejects points that lie on edges or ridges, whereas the medical imagery being analyzed comprised of regions that were edge/ridge in nature (bones etc).

Aman et al. (2010) modified the 3D SIFT methodology of Cheung and Hamarneh (2007) for content-based image retrieval in colonic polyp diagnosis. Rather than use image gradients to derive the descriptor histograms, they formed histograms from neighbourhood shape index (Koenderink and van Doorn, 1992). This was done to overcome the poor performance in regards of rotation - a flaw with the approach of Cheung and Hamarneh (2007) in implementing a 3D extension to SIFT. In this way they deviate significantly from the the original 2D implementation of SIFT (Lowe, 2004). They reported good results using synthetic data but noticeably poorer results when using real CT imagery - true-positive rate of 61% and a false-positive rate of 35% both with margins of error in excess of 20%.

It is notable that the primary focus of the prior work has been in action recognition, medical registration or medical panoramas as opposed to explicit object recognition. Only recently have attempts been made at object recognition (Aman et al., 2010) and in that case the SIFT methodology was significantly altered from the spirit of the original approach. There is no evidence of the SIFT descriptor being extended into 3D in a manner akin to that of the original application (Lowe, 2004).

Given the importance of the SIFT method in recognition tasks its use has been investigated in our work. The details of our own 3D extension of SIFT is discussed in Chapter 4. The work of Allaire et al. (2008); Aman et al. (2010); Dalvi et al. (2010); Ni et al. (2009); Riemenschneider et al. (2009); Niemeijer et al. (2009) was concurrent with the work outlined in this thesis.

2.2 Class recognition

Recognition of a known object has its limitations - we may wish to recognize a complete category or class of objects. For example, in the baggage-scanning envi-

ronment, it may be possible to identify one type of handgun but that is of limited use given the large number different handguns in existence. A solution that can cope with intra-class variation and thus identify any handgun from prior knowledge of the types of features or shape of handguns is required to address this problem.

2.2.1 Bag of words/features

The bag-of-words method for object class recognition has its origins in text analysis. From a given piece of text, a histogram of word occurrences can be built that can act as a descriptor. This descriptor is a compact representation of the text that has been used for retrieval applications ("find me an article like this"), identification of spam email and detection of plagiarism.

Sivic and Zisserman (2003) took the bag-of-words model to construct a 'visual word vocabulary' following analysis of video frames. SIFT descriptors were used on image regions located using both MSER and a Shape Adapted (SA) detector. The descriptors computed from each analyzed video frame are then clustered using K-means clustering (MacQueen, 1967) with $K \approx 6000$ for the MSER derived descriptors and $K \approx 10000$ for the SA descriptors. For a given frame, the descriptors are assigned to the closest cluster such that each frame is then described by a vector of visual word frequencies. A demonstration of object recognition is made that identifies video frames in which the specified object is seen.

Csurka et al. (2004) introduced the bag-of-keypoints method for recognition of objects by their class. Images of faces, buildings, trees, cars, telephones, bikes and books were used for training and testing. The SIFT methodology is applied to each image resulting in a number of interest point locations and subsequent descriptors. The approach taken was to use a fixed number of interest points, the locations of which in descriptor space are derived using K-means clustering ($K = 1000$) on the descriptors derived from a given set of training images. These clustered interest points form the dictionary for the bag-of-keypoints. A histogram is then built that shows which of the clustered descriptors are present in the image and this then acts as a the bag-of-keypoints for that image. Each image is now characterized by a histogram indicating the presence or absence of the range of keypoints in the visual word vocabulary. Classification using a linear support vector machine (SVM) and Naïve Bayes (Lewis, 1998) were performed using the histograms derived from training and testing sets of images. The SVM outperformed the Naïve Bayes classifier with an overall error rate of 15% (best performance: faces; worst performance: telephones).

An important aspect of the codebook formulation is the method chosen to as-

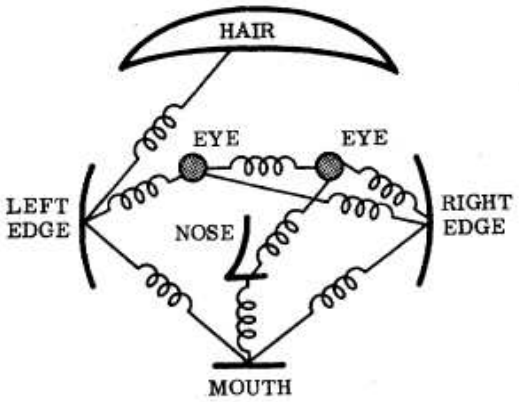


Figure 2.10: Modelling a face as a collection of rigid objects connected by "springs" (taken from Fischler and Elschlager, 1973)

sign each keypoint descriptor to one (or more) codebook entries. Hard assignment methods (Sivic and Zisserman, 2003; Jurie and Triggs, 2005; Nowak et al., 2006) assume that a given descriptor can only belong to one cluster centre. The assignment may be weighted (Sivic and Zisserman, 2003) in order to adjust the importance of that visual word in the codebook. Alternative approaches allow for some degree of uncertainty in the assignment by weighting each assignment to more than one cluster. So called 'soft' assignment methods have been shown to improve performance (van Gemert et al., 2010; Philbin et al., 2008) in overall recognition/categorization results.

2.2.2 Relational part models

The bag of words approach ignores any spatial or geometrical relationship between parts of an object in the recognition process. It seems logical to assume that better recognition performance would be achieved if this information were included as part of a recognition algorithm. We briefly give an overview of part model approaches which encompass this concept.

Fischler and Elschlager (1973) introduced the concept of modelling an object as a collection of rigid sub-parts conceptually connected by "springs". The springs allow the object some flexibility to recognition whilst also forming a recognition cost - the more deformation in the springs the less likely that an object has been found. An example given in Fischler and Elschlager (1973) is the modelling of a face and is shown in Figure 2.10. We see that the major facial features are modelled as rigid objects with deformation accounted for in their spatial relationship.

Felzenszwalb and Huttenlocher (2005) tackled estimation of human pose through the formation of an articulated model comprising 10 sub-parts. Figure 2.11 shows

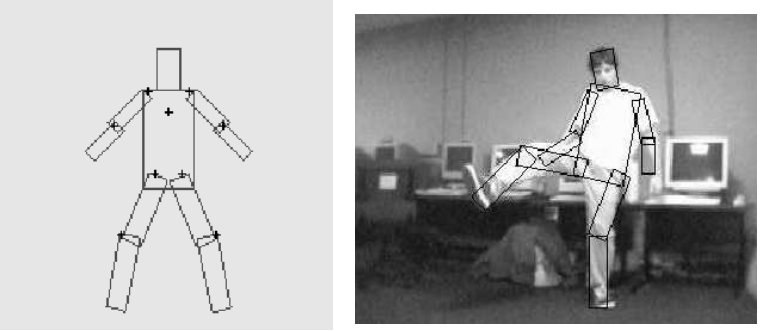


Figure 2.11: Articulated model of human body (left) and location in complex image (right) (taken from Felzenszwalb and Huttenlocher, 2005)

this model and an example image (taken from Felzenszwalb and Huttenlocher, 2005) matching the model to a human. The initial model did not specify the relationship between the parts but this was learnt from example images where the location and orientation of the various sub-parts has been specified. The work used binary images created through subtraction of the background prior to model fitting and pose estimation. No statistical results were presented.

Felzenszwalb et al. (2008, 2010) extended the work of Felzenszwalb and Huttenlocher (2005) to more general class recognition without the need for background subtraction, and achieved state of the art recognition for object classes in the PASCAL challenge dataset (Everingham et al., 2010). They model an object class using "visual grammars" - a hierarchical approach that defines the relationship between the object sub-parts. Rather than define the object model directly, it is learnt from example data. When presented with a previously unseen image the methodology forms two pyramid structures to account for variations in scale. The first pyramid is a conventional image pyramid. Features are calculated from each image in the pyramid in a dense grid in order to form a feature grid pyramid - the chosen feature method is closely related to the Histogram of Oriented Gradients (HoG) (Dalal and Triggs, 2005). Each location in the feature grid pyramid is analyzed to maximize a match metric that specifies the recognition of each sub-part and this is used to decide on the presence of an object. Figure 2.12 shows some examples of the detections made using this approach.

Felzenszwalb et al. (2010) note that "to obtain high performance using discriminative training it is often important to use large training sets." This has implications for our work as we are limited in the number of baggage scans available. Part model approaches have proven to produce good recognition performance in 2D imagery of natural scenes but can be complex in nature. It may be that an approach taking inspiration from the human vision system could, in time, surpass part models using

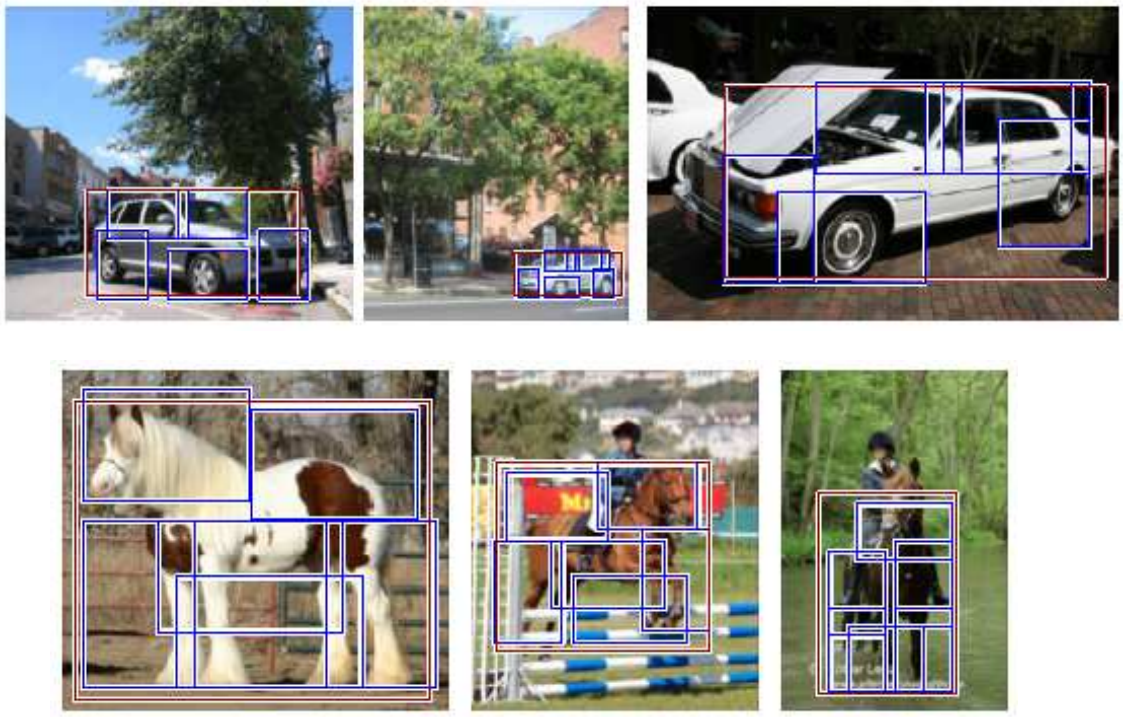


Figure 2.12: Recognition of cars and horses using part models (taken from Felzenszwalb et al., 2010). Blue boxes indicate sub-part recognition.

simpler processing structures.

2.3 Modelling the visual cortex

The descriptor-based methodologies (SIFT (Lowe, 2004), SURF (Bay et al., 2008) etc.) whilst claiming inspiration from biological vision are not direct models of mammalian vision systems. Biological vision has been an area of research interest for many years and computer modelling has recently yielded results that are of interest in the 3D-recognition task we are faced with.

The visual cortex is located at the back of the brain and is the region responsible for processing visual information arriving from the retina via the optic nerve and subsequent brain structure. Various mammalian brains (cat, rabbit, macaque monkey, spider monkey) have been studied as a means to derive functional models (Hubel and Wiesel, 1959, 1962, 1968). The visual cortex appears to be arranged in distinct sub-regions with one sub-region, called the primary visual cortex (V1), being the most studied area. It was discovered that V1 is hierarchical in structure with simple (S) and complex (C) neurons forming the basis of the hierarchy (Riesenhuber and Poggio, 1999). In the examination of V1 functionality it was discovered that some of the simple neurons in the V1 region respond to oriented bars and edges

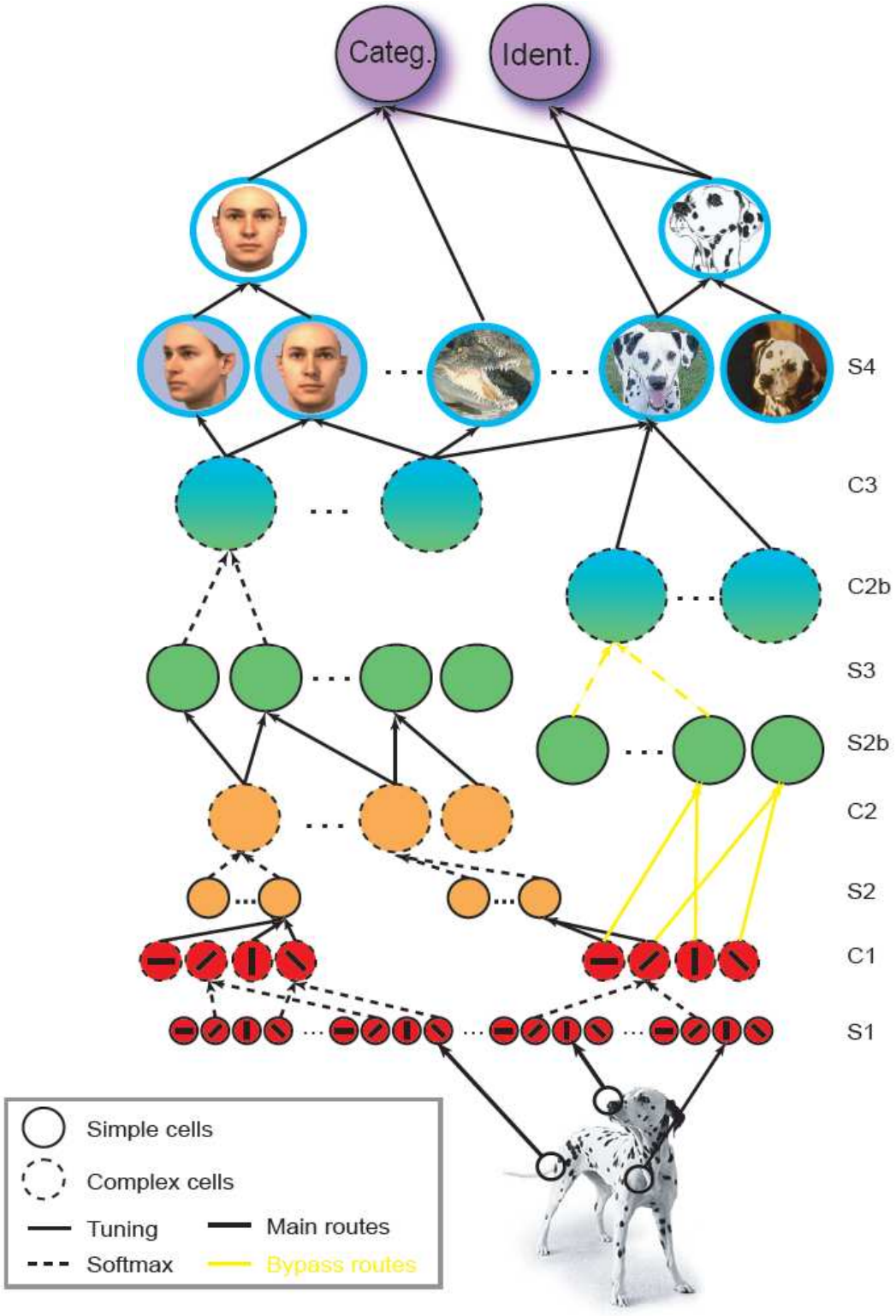


Figure 2.13: Visual cortex hierarchy proposed by Serre et al. (2005a)

(Hubel and Wiesel, 1959, 1962, 1968). This has led to the use of Gabor filters of varying orientation being used as the first processing stage in software models (Serre et al., 2005b; Mutch and Lowe, 2008; Jhuang et al., 2007).

Serre et al. (2005a) proposed the hierarchy shown in Figure 2.13 showing alternating simple and complex cells. In this model the V1 region is modelled by layers S1 and C1. Beyond S1 and C1 are modelled higher regions in the visual cortex such as V4 and the inferotemporal cortex.

Moving up through the hierarchy it has been found that the cortex response is increasingly invariant to object transformations (scale, position) whilst also becoming focused on more specific features (relevant to objects of interest). These processes have been modelled through the use of "max-pooling" operations (Riesenhuber and Poggio, 1999) and subsequent comparison with learnt salient patches.

At present the V1 region is primarily viewed as a feed-forward path from the reception of an image at the retina to the higher layers in the visual cortex (Riesenhuber and Poggio, 1999; Serre et al., 2005b) where the received image is processed by the simple and complex computational hierarchy without any feedback paths being present.

In the work of Serre et al. (2005b) a comparison was made between the visual cortex model and the SIFT descriptor (Lowe, 2004) for generalized object recognition using a support vector machine for classification on the Caltech 101 dataset (Fei-Fei et al., 2007). In this implementation a bank of Gabor filters are used in the S1 layer comprising 8 bands and four orientations (0°, 45°, 90°, 135°) with progressively larger wavelengths ($\lambda = 3.5$ to $\lambda = 22.8$ pixels) to account for feature-scale selectivity. Figure 2.14 shows examples of the imagery used in the recognition task. The work reported results for which the false-positive rate equalled the false-negative rate with recognition rates in excess of 93% achieved using linear SVM or Ada Boost classification. Classification of 'cars' was near perfect (above 99.7%). A comparison was also made between the visual cortex model and the use of SIFT descriptors in the classification process which appeared to show that the visual cortex approach outperformed the SIFT method by some margin, though it should be noted that this comparison was just on the descriptors - a comparison of SIFT with the bag-of-features model may have been a more reasonable experiment. An interesting aspect of this work is that with the dataset used by the authors of (Serre et al., 2005b) high recognition rates can be achieved using comparatively few training examples - 30 training images can produce a true-positive rate in excess of 90%.

Jhuang et al. (2007) applied a small 3D extension of the visual cortex models of Serre et al. (2005b) and Mutch and Lowe (2006) for the purpose of action recog-

Figure 2.14: Example cars and faces used in recognition task (taken from Serre et al., 2005b)

nition. The incorporation of temporal information requires interpretation in three dimensions yet a spatio-temporal volume is not formed in their work. Three filtering approaches are evaluated for the S1 layer with one comprising a 3D spatio-temporal filter tuned to 4 spatial directions (0°, 90°, 180°, 270°) and 2 motion speeds (3 and 6 pixels/frame) calculated over 9 pixels × 9 pixels × 9 frames. Thus the output for each pixel is a vector comprising 8 elements where each element indicates a particular motion speed and direction in the video frame. This interpretation means that the output of the S1 layer is 2D vector image, as used in Serre et al. (2005b) and Mutch and Lowe (2008), with 8 elements rather than 4 per pixel. From this point on the work follows that of Serre et al. (2005b) with the implementation of the 2D C1 layer and onwards to a linear SVM classifier. The addition of a new simple/complex layer (S3, C3) is made with the intention of introducing temporal invariance to the model (spatial invariance is achieved in the C2 layer). A number of human actions were analyzed with typical recognition rates of $\approx 90\%$.

The work of Serre et al. (2005b) was extended by Mutch and Lowe (2008) in a number of ways. Instead of the applying Gabor filters of increasing size to a single image, a scale-space pyramid was constructed that allowed the same Gabor filters to be applied at all levels and achieve the same results with much less processing. They then improved classification through a number of changes to the processing function in the model hierarchy. Firstly, sparsification of the S2 layer input unit was made in a bid to mimic likely neuron input weighting behaviour in the visual cortex. Implementing sparsity removes some useful information: this is compensated for by

increasing the number of spatial orientations from 4 to 12. Inhibiting relatively weak S1/C1 outputs within a patch was also implemented using a simple threshold, again to mimic neuron behaviour. Limiting the position and scale invariance in C2 was achieved by limiting the region (in position and scale) where a given S2 feature can be located. Finally, the classification features undergo a selection process rather than being just randomly selected from the training images (as in Serre et al., 2005b) in order to choose features that are significant to the recognition task. Final performance on the Caltech 101 dataset showed an improvement over the original approach (Serre et al., 2005b) with 56% recognition rate when using 30 training images (was 42% in Serre et al., 2005b). This level of performance is not as good as can be achieved using other techniques. For example, Bosch et al. (2007) achieved about 80% recognition on the same dataset in part through selection of regions of interest prior to classification.

Walther et al. (2002) used saliency maps to modulate the S2 layer rather than explicitly extract regions of interest in an attempt to improve recognition. One class of object was used (twisted paper clips) and it was shown that a small amount of modulation to the S2 layer increases recognition performance. No experiments were performed on scenes of natural imagery.

Huang et al. (2011) used the model of Serre et al. (2007) for the classification of 4 types of natural scenes (corridor, office, garden, road). They modified the patch selection phase from the purely random approach in Serre et al. (2007) to one that samples from salient regions. The classification results showed improvement when using the salient patch selection process.

We believe that the standard model has not been fully extended into 3D for the purpose of object recognition.

2.4 Baggage security applications

There is little published work in the area of automatic recognition of items in scanned baggage, from both X-ray and CT. Below is a review of the available published literature.

Bi et al. (2008) used a CT scanner to detect a handgun in baggage. The work did not involve processing the 3D data directly - the problem was reduced to searching for the characteristic cross-section that the handgun presents, and appeared preliminary in nature. No results are presented in respect of detection performance. Further work by the same author (Bi et al., 2009) presented a methodology for the detection of planar materials within CT-baggage imagery using a 3D extension to

the Hough transform (Ballard, 1981). The reported work concentrated on implementing the algorithm on a Graphics Processing Unit (GPU) and again no results were presented on the performance of the detection method itself.

Chan et al. (2009) investigated the use of the SIFT methodology on kinetic depth effect (KDE) X-ray imagery (Evans, 2003) - a method of forming a 3D image from a sequence of 2D X-ray images. In this regard the work was not aimed at automatic detection of threat items, rather it was aimed at improving the quality of the imagery presented to human operators.

Nercessian et al. (2008) investigated 2D X-ray imagery of luggage for the detection of handguns. The method uses edges detection to characterize handgun features but only deals with handguns in a fixed orientation. A small dataset was used (40 images with handguns, 400 images clutter). Two simple examples of handgun detection were shown but no statistical results were presented.

Baştan et al. (2011) recently applied the bag-of-words model to colour 2D dual energy X-ray images of baggage items to detect handguns and are the first to report detection results. Dual energy X-ray scanners illuminate the baggage item with a high and low power X-ray beam from which an estimate can be made of the material type present at each pixel location. This material type image is coloured to aid human operatives recognize objects. Figure 2.15 shows such a scan (taken from Baştan et al., 2011) where we can see three images: high power gray-scale; low power gray-scale; material type colour derived from the gray-scale images. Investigation of a variety of interest point detectors (DoG, Hessian-Laplace, Harris, FAST, STAR) coupled with three descriptors (SIFT, SURF, BRIEF) was made. Whole baggage items were considered rather than cropping threat items which raises the complexity of the recognition task. Vector quantization using soft assignment produced the best results using a support vector machine with an intersection kernel (Maji et al., 2008). In the classification of baggage containing handguns they reported that the method does not work well in isolation but results can be improved using the extra information available from the colour image (indicating material type).

Megherbi et al. (2010) investigated detection of bottles in CT volumetric data. Baggage items are segmented from CT volume and analyzed using two approaches. The first approach is to form a descriptor by analysis of the surface of the extracted item. A normalized histogram of shape index (Koenderink and van Doorn, 1992) is formed as the descriptor from this approach. The second method uses Zernike descriptors (Novotni and Klein, 2004) which are rotation invariant and produce a finite vector description of the voxel-based object derived from a set of orthogonal basis functions. From a small dataset (79 volumes for training, 126 for testing)

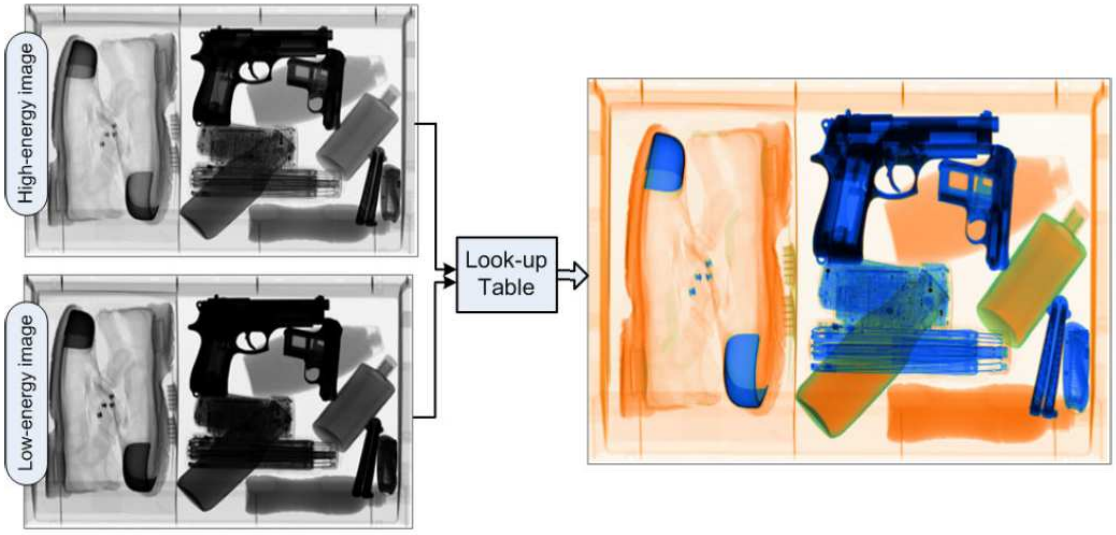


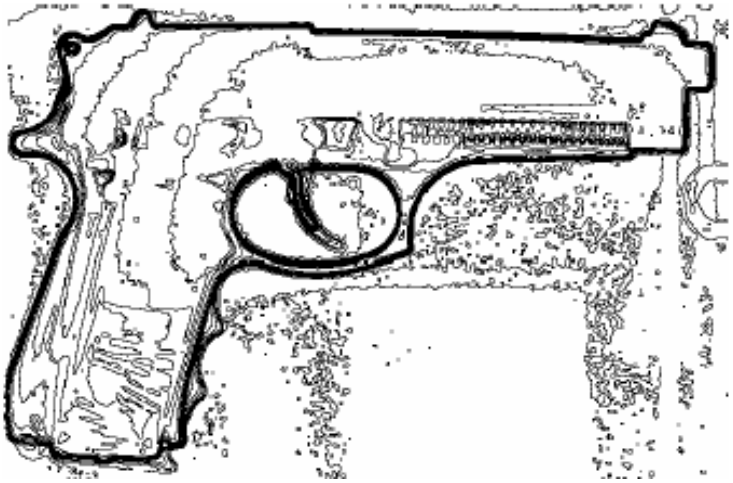
Figure 2.15: Dual energy X-ray producing material type image (taken from Baştan et al., 2011)

classification results in excess of 98.0% accuracy were obtained using the histogram of shape index.

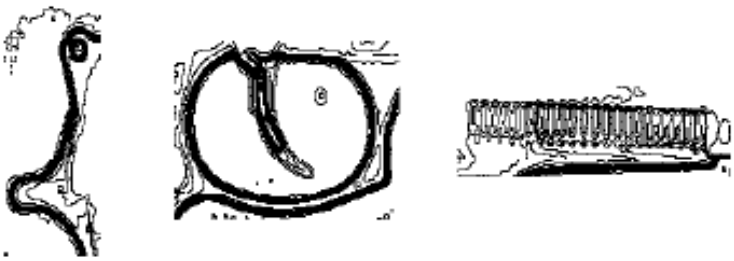
Oertel and Bock (2006) tackled detection of handguns in 2D gray-scale X-ray baggage imagery. The research is an example of specific item recognition: one type of handgun was characterized using the distinguishing features of its trigger, hammer and spring. These features can be seen in Figure 2.16. Regions of interest are created for each pixel on an edge contour and a descriptor is constructed from the distribution of white and black pixels in the local neighbourhood. It is unclear if is this is rotation invariant in the horizontal plane and the method is certainly not invariant if the weapon is rotated out of this plane. A small dataset was used (40 X-ray images: 30 for training, 10 for testing) and no quantitative results were produced.

Gesick et al. (2009) were also interested in detecting handguns from 2D X-ray imagery. Edges were extracted from the imagery and the handgun trigger guard was searched for in a similar fashion to (Oertel and Bock, 2006). The method was not rotation invariant and no quantitative results were produced.

The majority of articles covering threat recognition in volumetric data are not generalized class recognition approaches but single instance recognition methods for which performance is generally poor. A significant amount of prior work only deals with 2D imagery and additionally does not use rotation invariance approaches. Given the lack of articles covering threat recognition in volumetric data, we will now broaden this review to include recognition techniques that are used in medical image analysis.



(a) Reference item: 9mm Colt Beretta



(b) Characteristic feature contours: hammer, trigger, recoil spring

Figure 2.16: 2D X-ray handgun recognition (taken from Oertel and Bock, 2006)

2.5 Medical scanning

Given the lack of published research on threat detection in CT-baggage imagery it seems prudent to study research in the medical image analysis community. The use of CT and MRI imagery in this area has spurred research into automatic computer vision recognition of items of interest to aid diagnosis in a number of medical conditions (see Doi, 2007, for a review). An important distinction here comes from the use of the word "aided". The computers are not making the diagnosis but assisting an expert user by pointing out abnormal regions in images. It is unlikely that human experts will be removed from the diagnosis completely for ethical reasons and it seems a reasonable assumption that this will be the same situation in baggage scanning.

As we have seen (Section 2.1.5.2) computer vision techniques have been used to perform registration in 3D for CT, MRI and ultrasound imagery. Volumetric stitching has also been addressed for ultrasound imagery. The use of computer vision approaches for recognition of objects in medical imagery is of great interest. For example, one area of medical analysis is the recognition of colonic polyps - abnormal tissue growth which can be pre-malignant and must be removed. Recognition of polyps using computer vision techniques appears to take two primary forms: direct interpretation from the voxel volumetric data or shape analysis of the surface the colon wall.

Suzuki et al. (2008) looked for colonic polyps that had been missed by expert users on 3D CT data. Given that experts had not identified the polyps we can assume that they are difficult to recognize. The approach taken used example subvolumes containing either a polyp or non-polyp as training data for an Artificial Neural Network (ANN). The overall result was a detection rate of 71.4% on those polyps that had been missed by the expert users with the number of false positives being declared as "relatively small" (approximately 5 per patient).

Yoshida et al. (2002) used surface shape to detect polyps in CT imagery of the colon. The colon wall is extracted as an isosurface and the volumetric shape index and volumetric curvature are calculated at every point. The shape index indicates the type of shape that is present (cup/ridge/polyp etc) whereas the curvature is a measure of how flat or sharp a potential polyp shape is. Figure 2.17 shows example shapes including a polyp taken from the isosurface. Use of these measures showed good detection of polyps but also a relatively high false-positive rate. This was reduced through examination of the local gradient around the candidate polyp feature - the gradients will tend to point in towards a polyp. Final results showed near-perfect detection with approximately two false positives per patient.

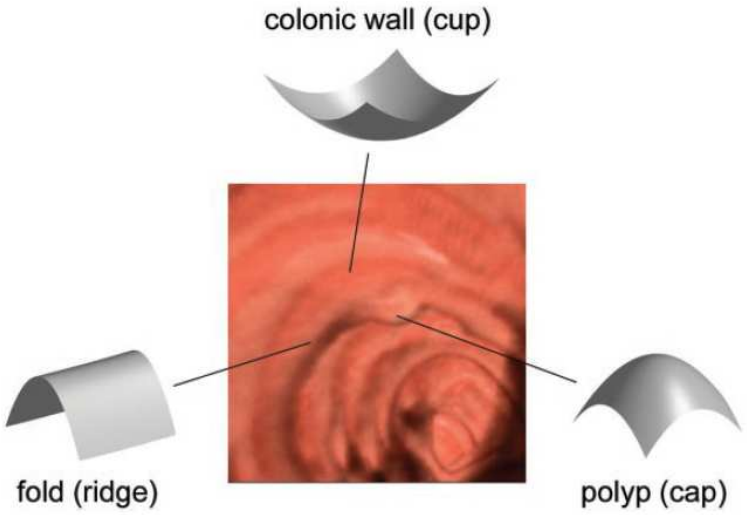


Figure 2.17: Surface shape characteristics can be used to detect polyps (taken from Yoshida et al., 2002)

For both colonic polyp CAD systems (Suzuki et al., 2008; Yoshida et al., 2002) the ideal result would be 100% detection of genuine polyps with a small number of false positives. Expert radiologists would verify all detections with the aim being that their workload is reduced as they no longer have to locate candidate polyps within the CT imagery.

Tu et al. (2006) used 3D Haar filters to search for colonic polyps, taking inspiration from successful 2D face detection methods (Viola and Jones, 2001). This approach used the voxel data directly - no surfaces were extracted. Rotation invariance was tackled using a hemispherical polyp template that can be used to crudely align the candidate region in both direction and scale. Training and classification was made using a probabilistic boosting tree (Tu, 2005). Recognition rates of 98% for small polyps and 84% for large polyps were reported with 3 false positives per scan. The authors note that there is not a common dataset for researchers to use to evaluate their polyp-detection algorithms.

Searching for lung nodules is another area of research (Antonelli et al., 2011; Liu et al., 2009; Sluimer et al., 2006) where computer vision techniques are being applied. Statistical measures of shape and texture seem to be the approaches taken rather than a bag-of-words approach which can be explained by the non-rigid nature and high variability in the structures of the human body.

2.6 Summary

There are several computer vision techniques that could be applied to the recognition of objects within 3D CT-baggage imagery ranging from conventional 2D techniques through to applications in the 3D medical domain. It is apparent that little prior work has been carried out in the specific area of object recognition within CT-baggage imagery and this reveals a number of opportunities to extend knowledge:

- The application of 3D SIFT for explicit specific object recognition in 3D imagery. This would allow an examination of the performance of the 3D SIFT descriptor within this imaging paradigm and the creation of a recognition system for baggage items.

- Class recognition of baggage items in 3D CT using interest points implemented through either a bag-of-features or part model based approach. The ability to recognize a class of object would be the logical approach to take in a deployed recognition system.

- Extension of the Visual Cortex Standard Model to 3D for recognition in 3D imagery. The development of a 3D extension to a biologically inspired 2D technique would provide a direct comparison to the interest point class recognition approach.

Our approach begins with development of a 3D SIFT implementation for investigation of specific-instance recognition akin to that of (Lowe, 1999, 2004). With confidence in a 3D SIFT implementation we can extend into class recognition (Sivic and Zisserman, 2003; Csurka et al., 2004) within the CT data. Extending the visual cortex model (Mutch and Lowe, 2008) will then contrast the interest point based class recognition.

The quality of the CT data used for this study is poor - numerous imaging artefacts and noise are present. An appreciation of the CT data and its associated artefacts is required before we begin experimentation with 3D SIFT, and this is presented in the following chapter.

Chapter 3

Source data

Volumetric 3D data used for this research is derived from a CT scanner specifically designed for the task of baggage-content examination. The imagery is generally poor in nature, suffering from numerous artefacts and noise, and differs significantly from medical imagery of the same genre. As we have seen (Chapter 2) there is little prior work on recognition within CT-baggage imagery. Consequently we will now discuss the source of the imagery and the processing that is applied prior to application of subsequent recognition algorithms outlined in this thesis in order to allow appreciation of the noise/quality problems to the reader.

3.1 CT scanner

3.1.1 Overview

Volumetric CT imagery was obtained using a CT-80 baggage scanner manufactured by Reveal Imaging Inc. and shown in Figure 3.1. The primary focus of this scanner is the detection of explosive and organic materials using dual energy CT techniques (Ying et al., 2007) - high and low energy X-rays are used to probe the baggage item and the differing response to each is used to estimate the material type within. A consequence of the application of the scanner to commercial baggage scanning is the speed of item transit. This an important aspect of the design and leads to compromises in image quality in terms of both resolution and noise.

Figure 3.2 shows a cross-section through the scanner. This shows an X-ray source and a detector bank on the opposite side. The X-ray source and detector are rotated around the central axis with the detector responses being recorded for each angular position, θ. This results in a 2D detector image called a sinogram, an example of which is shown in Figure 3.3. Conversion of the sinogram image to a form which

Figure 3.1: Reveal Imaging CT-80 baggage scanner

shows density distribution throughout the scanned object requires an inverse Radon transform (Kak and Slaney, 1988). Figure 3.3 shows an example image for one slice where we can see the detector responses as the X-ray source and detectors are rotated. This shows characteristic wave-like features that result from the presence of localized rigid objects in the baggage item (in this case a pistol and batteries). Application of an inverse Radon transform to this sinogram image results in the conventional CT slice image, as shown in Figure 3.3b.

The CT scanner comprises a conventional belt system to transport a baggage item from entrance to the exit. X-ray absorbent curtains are placed at the entrance and exit to prevent harmful radiation emissions. The belt system can be operated in two modes: a stepping mode, where the bag is moved in a series of steps and held stationary for each slice; a faster constant velocity mode, where the bag is moved at a constant speed through the machine whilst scanning takes place. Figure 3.4 shows the effective scanning arrangement for the stepping and constant velocity modes of operation. It can be seen that the stepping mode results in a set of well defined vertical imaging slices: the belt is stationary when a scan is made yielding a 'static' scan. For the constant velocity mode, the continuous motion of the baggage item is equivalent to a helical motion of the scanning hardware. The helical scan must be converted into a set of slices which results in additional imaging artefacts (see Section 3.1.3.2).

The imagery produced by the scanner is a collection of 2D image slices that can be formed into a 3D volume. The scanner automatically senses the baggage extent in the axial direction and limits the number of output slices to those that cover the scanned object. Two images are recorded for each slice: high power and low power. We choose the high power slices as they are less susceptible to imaging artefacts and noise (see Section 3.1.3).

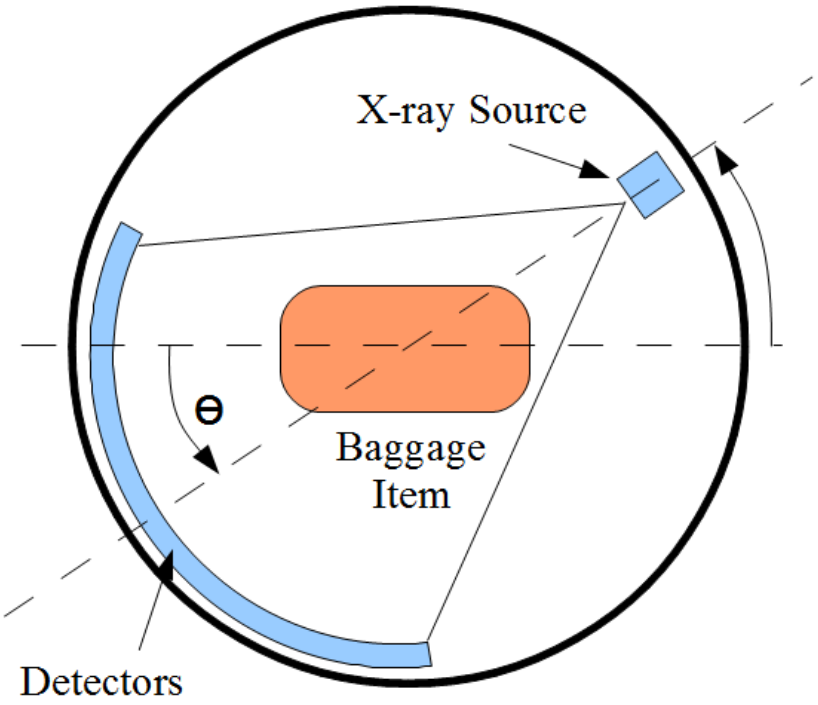


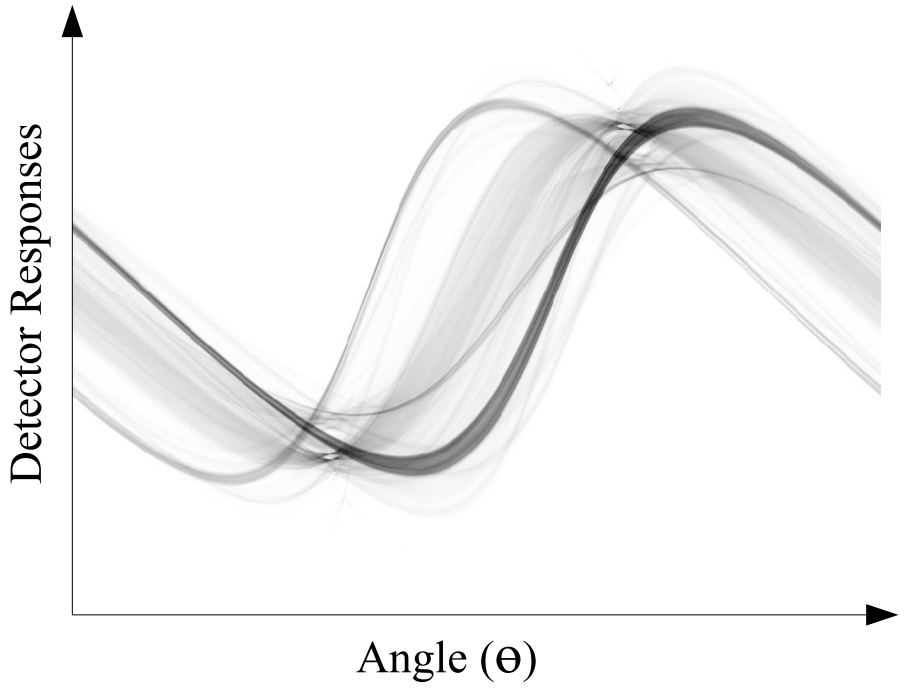
Figure 3.2: Cross-section through CT scanner

3.1.2 Resolution

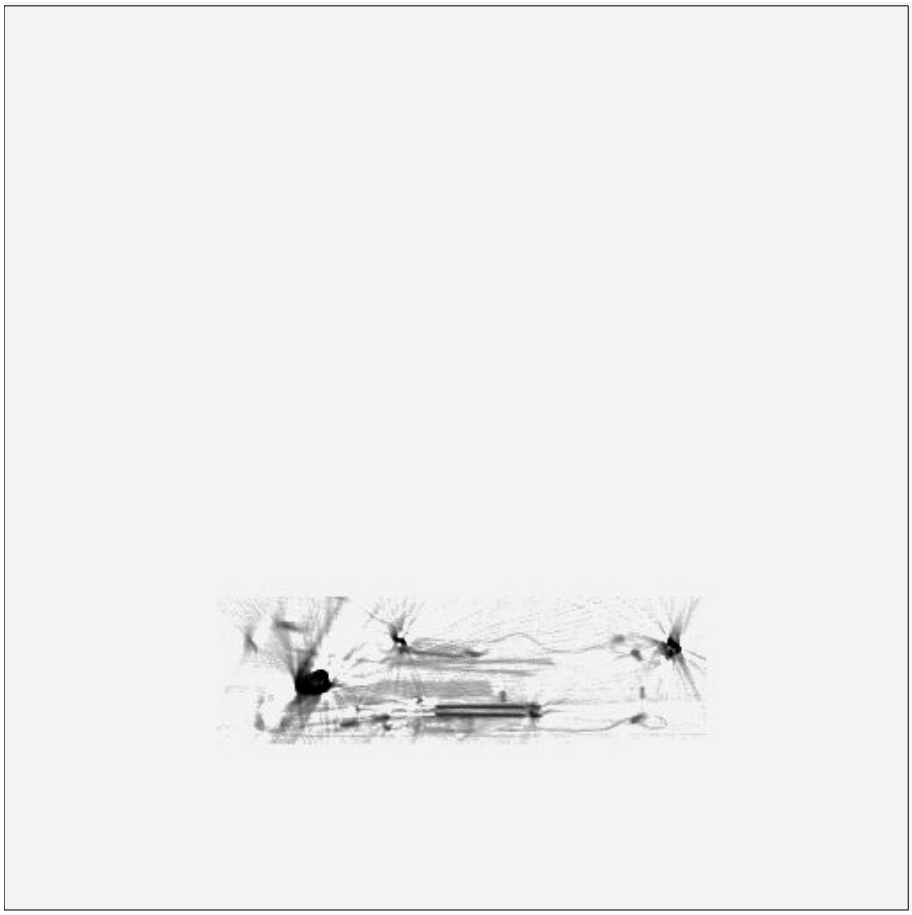
Each captured slice from a baggage scan is 512×512 pixels in size with each pixel representing a physical size of $1.56mm \times 1.61mm$. Each slice is spaced $5mm$ - a hard limitation imposed by the X-ray imaging capabilities of the CT-80 scanner. Achieving a $5mm$-slice spacing requires that the machine is operated in a mode which increases the time taken to scan a typical baggage item from $\backsim 10s$ (achieved with $16.5mm$-slice spacing and helical scan) to $\backsim 200s$. Initial work carried out using the default slice spacing ($16.5mm$) indicated that recognition of items would be extremely difficult: such a large slice spacing fails to image the baggage item completely, effectively leaving $11.5mm$ gaps between slices. Such a large spacing leads to a failure to accurately image objects within the baggage.

3.1.3 Imaging artefacts

A common problem in medical CT images is the presence of imaging artefacts (defined below). Barrett and Keat (2004) give an introduction to this topic and highlight the main artefacts that occur in medical images. Wang and Vannier (1994) discuss the "stair-step" artefact (see below) in detail. Some of these artefacts have been noticed in the CT images of baggage obtained for our research and these will

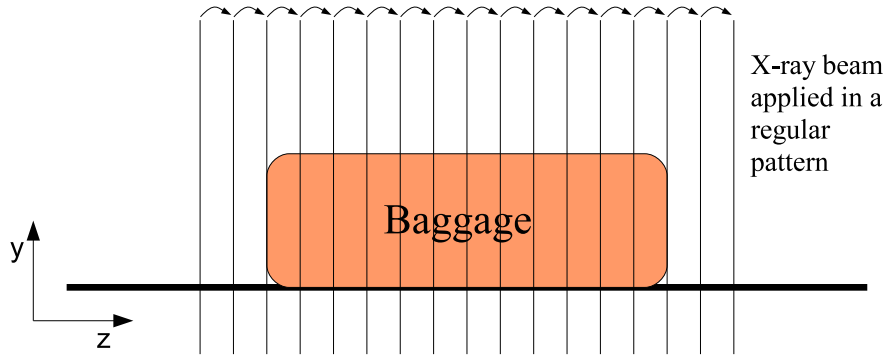


(a) Sinogram: detector responses as X-ray source and detectors are rotated

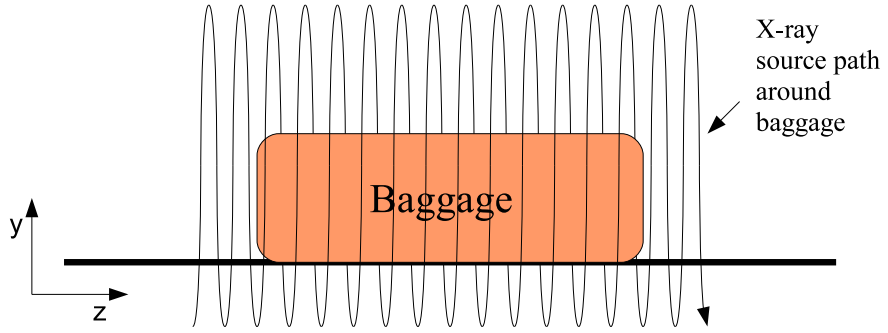


(b) Output of inverse Radon transform using data from Figure 3.3a as input

Figure 3.3: Use of inverse Radon transform to produce final slice image



(a) Stepping mode: static scan



(b) Constant velocity mode: helical scan

Figure 3.4: CT scanning modes: static or helical scans

now be discussed further as they have an impact on the performance of the recognition task we are addressing.

3.1.3.1 Beam hardening: streaks and shadows from metallic objects

Figure 3.5 shows a single CT slice from a bag that contains a metallic object (handgun) and a number of other clutter items (shoe, paper file). Large amounts of metal in the baggage lead to imperfect measurement in the CT scanner detectors, and hence the sinogram image. Imperfection in the sinogram image will lead to reconstruction errors in the final slice image that manifest themselves in two forms: streaks and shadows. Shadows are regions in the density image where failings in the original sinogram image have the effect of removing items from the presented image slice. Streaks appear as lines with apparent density that radiate in the xy plane. Both streaks and shadows can be seen in Figure 3.5. Given that the errors are in the xy plane, they will vary depending on the source object orientation in the scanner, and are not fixed for each object. They cannot be used as a method of identification.

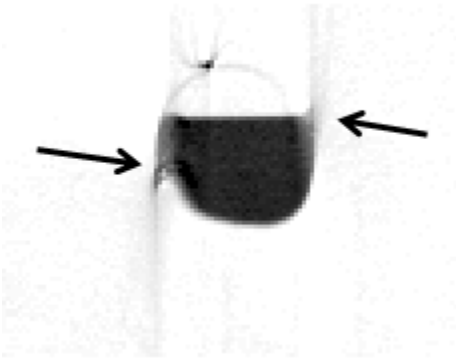


Figure 3.6: Example helical-scan artefact

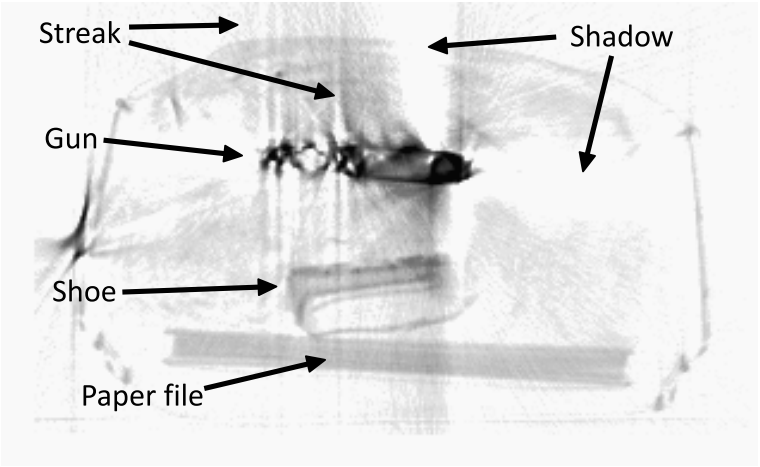


Figure 3.5: CT image streak and shadow artefacts caused by presence of metallic object (handgun)

3.1.3.2 Helical-scan artefacts

In normal operation the machine keeps the baggage item under constant motion as the X-ray beam is scanned across it. This gives rise to a helical scan which results in artefacts caused by errors in the reconstruction of each slice. Figure 3.6 shows an example of a helical-scan artefact. This shows the cross section of a bottle containing water. In this case we can see artefacts created either side of the bottle. This is due to errors in the creation of the final image slice from the helical-scan data.

For our work we operate the CT-80 in a mode that performs 'static' scans - the baggage item is scanned in a series of small steps thus eliminating the helical-scan artefacts.

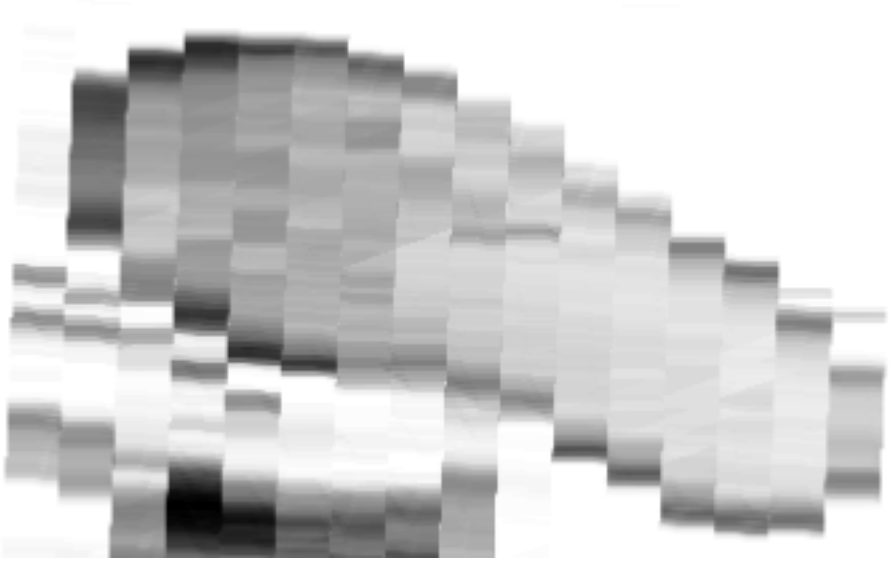


Figure 3.7: Stair-step artefact on scanned shoe

3.1.3.3 Stair-step artefacts

When the CT images are combined to form a volume, further artefacts are apparent. If the slice spacing is large compared to the X-ray beam width, then objects that are sampled will appear to have 'stair-step' edges. This artefact is, in effect, a basic consequence of the spatial sampling frequency: small slice spacing will reduce the magnitude of this effect. Figure 3.7 clearly shows the characteristics of the 'stair-step' artefact for a scanned shoe.

3.1.3.4 CT artefact correction

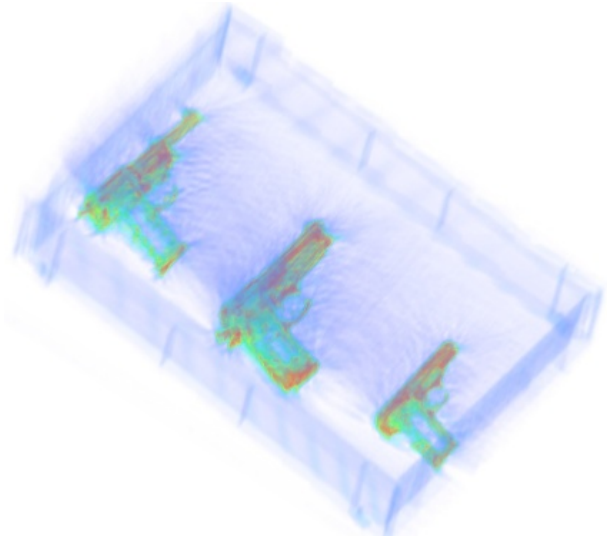
The removal of CT artefacts is of great interest to the medical-imaging community as artefacts can destroy key features of clinical interest thus risking misdiagnosis. Methods exist to correct artefacts due to metallic objects (Zhao et al., 2000; Wang et al., 1996) but these are applied on the raw CT X-ray data before the CT image slices are produced. It has not been possible to pre-correct for metallic-object artefacts within the imagery used in this project - reduction of these artefacts within the overall processing pipeline is identified as an area for future work.

3.2 Acquisition methodology

In order to obtain the highest quality imagery the CT-80 scanner was configured to use a $5mm$ slice spacing with helical scanning disabled.

A number of target items were used during the research. 'Clean' scans of these

(a) Tray of guns



(b) CT scan of tray

Figure 3.8: Example reference items being scanned

items, with little clutter, were taken to provide reference data for the object recognition algorithms. Figure 3.8a shows example reference items material (pistols) being scanned in trays with the associated CT imagery shown in Figure 3.8b. These items were subsequently secreted in baggage items and scanned.

Baggage items of varying types which contained various degrees of 'clutter' items were scanned. This included both hand luggage and hold baggage. Some examples are shown in Figure 3.9. Target items were inserted into the baggage items so that a database of baggage items with and without targets was obtained.

We also scanned some reference items in containers containing foam inserts, as shown in Figure 3.10. This raises the object away from the container walls to allow

(a) (b)

(c) (d)

(e) (f)

Figure 3.9: Example baggage

Figure 3.10: Container with foam inserts

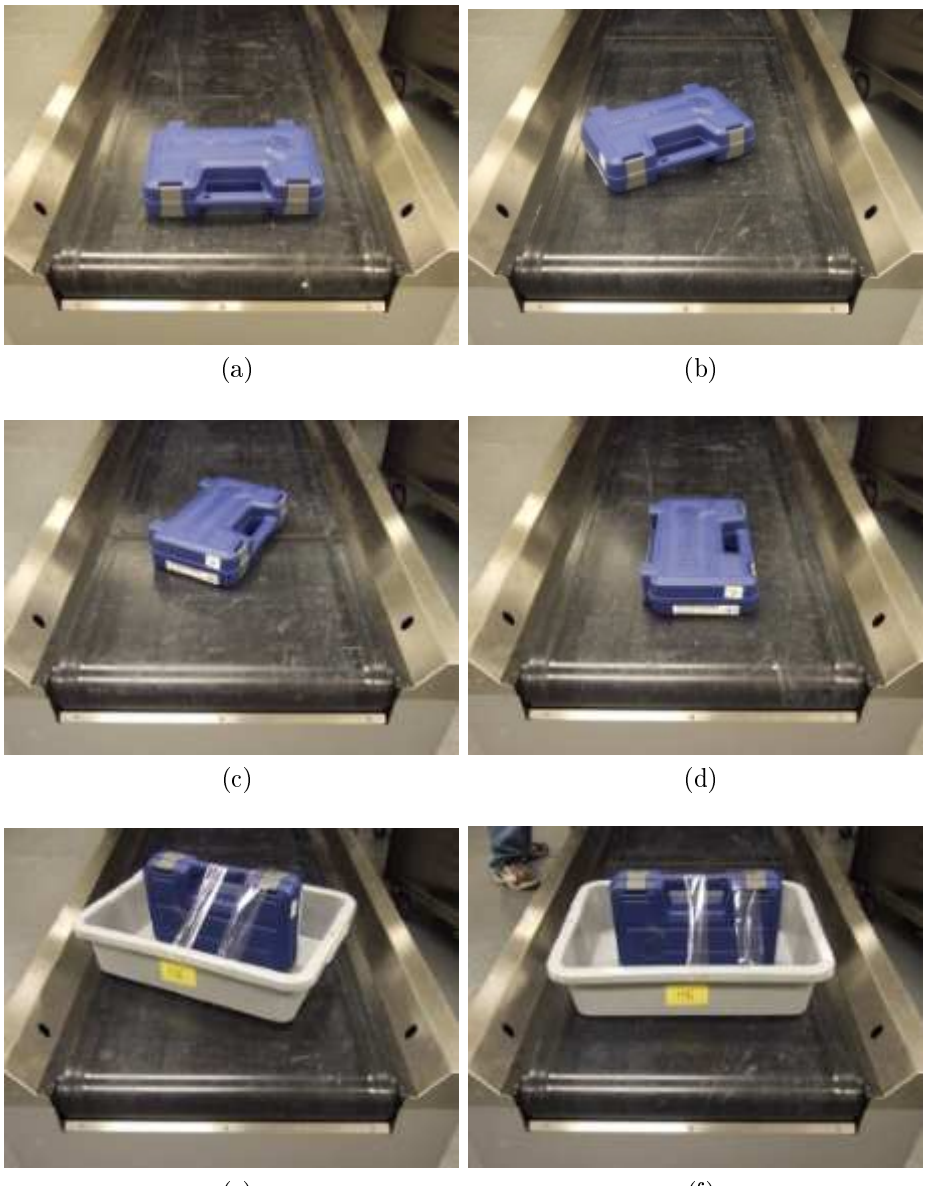
investigation remote from any clutter as well as allowing investigation of rotation on the scanning results by careful placement of the container prior to entry to the scanner, as shown in Figure 3.11. Precise placement of the item within the scanner was not possible as the scanner conveyer belt and radiation curtains will alter the container position as the the item enters the machine.

3.2.1 Target items

Figure 3.12 shows some of the target items that were scanned. It is important to recognize that the colours of the CT image of an object may not bear much resemblance to a normal camera image. For example, Figure 3.13 shows the Glock 26 handgun from Figure 3.12f in the CT domain. This pistol has a metal barrel but a plastic grip. The plastic grip has a much lower density than the barrel and consequently its appearance in the CT domain is considerably different its the visual appearance.

3.2.2 Clutter items

Baggage contents were packed with materials to represent clothing. Clothing has a low CT density and helps to separate more solid items in 3D space. Clutter items were added to the baggage such as those shown in Figure 3.14. The aim of the clutter to is provide baggage imagery that is comparable to that encountered within transport infrastructure. A lot of clutter is low density (clothes, books, etc) while high density items include any metallic items (belt buckles, batteries, etc).



(a) (b) (c) (d) (e) (f)

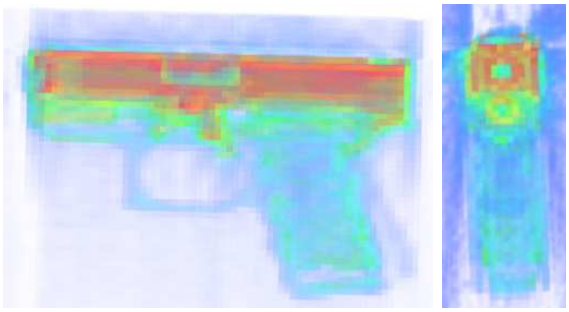
Figure 3.11: Differing orientation prior to scanner

(a) Smith & Wesson Magnum (b) Browning 7mm

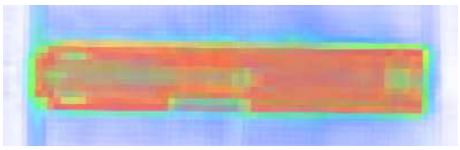
(c) Walther PPK (d) Bruni

(e) Glock 19 (f) Glock 26

Figure 3.12: Example handgun target items



(a) Side view (b) Front view



(c) Top view

Figure 3.13: Glock 26 under CT

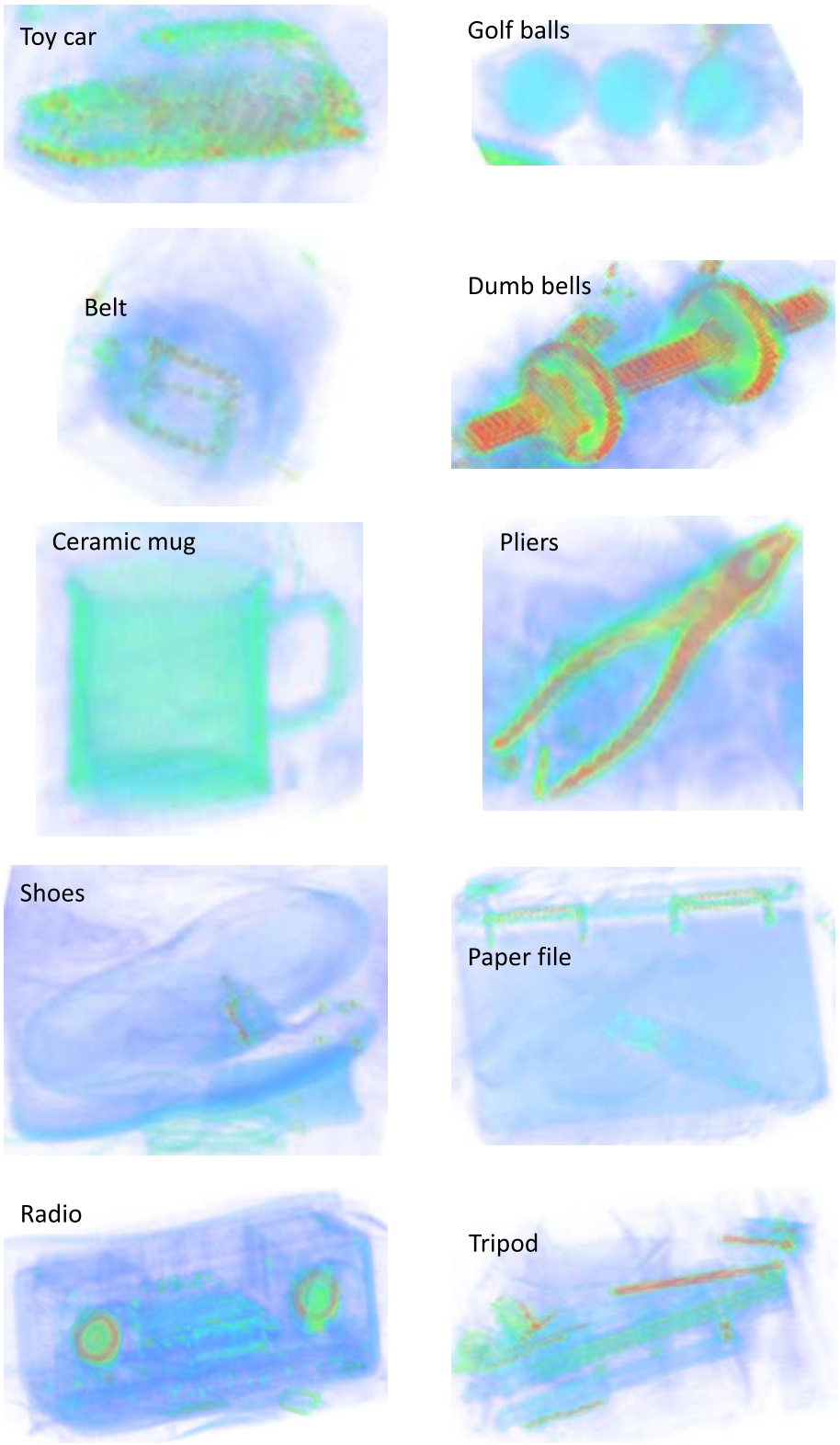


Figure 3.14: Example clutter items

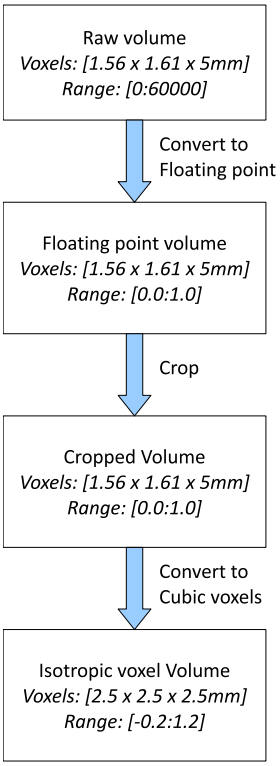


Figure 3.15: Cropping and resampling

3.3 Data representation and processing

The density at each pixel in an image slice is represented as in integer in the range [0, 60000] and is calibrated so that air has a density of 0 and water a density of 10000. When first analyzing the data we change to a floating point format and normalize such that air has a density of 0.0, water a density of 0.167 with a full scale reading of 1.0. This step eases subsequent algorithmic development as we are no longer concerned with accumulation overflows.

The anisotropic volume data, with voxels of $1.56 \times 1.61 \times 5mm$, are cropped in xy to leave the complete baggage item plus a small margin in order to reduce the amount of data stored and decrease the amount of subsequent processing. This cropped volume is then resampled using cubic interpolation to form a volume with isotropic voxels of $2.5 \times 2.5 \times 2.5mm$ as shown in Figure 3.15. This resolution was felt to be a reasonable compromise between resolution, interpolation error, processing time and storage and was chosen in order to simplify the algorithm. The cubic interpolation does result in the data value range increasing beyond [0.0, 1.0] but was not considered to be a concern - a value above 1.0 can be regarded as "more dense" and a value below 0.0 as "less dense".

3.4 Sub-volume generation

For some experiments we require target items to be cropped from the baggage items in order to form a training set. Similarly we require smaller sub-volumes of clutter for a comparable dataset. Extraction of items of interest was performed such that a $5cm$ (20 voxels) margin was maintained around the item.

Clutter sub-volumes were obtained by splitting whole clutter volumes into sections: each dimension (xyz) was subdivided to leave the number of voxels in the range [64, 128]. This range was chosen as it is similar to the handgun and bottle sub-volume dimensions used during the experiments; thus the clutter sub-volumes are a similar size to the target sub-volumes.

Examples of handgun sub-volumes are shown in Figure 3.16. Note that the volumes have been rotated for best visibility of the target item in the 2D projections shown.

Examples of clutter bag sub-volumes are shown in Figure 3.17.

3.5 Summary of data

Access to the CT scanning machinery during this work was limited to a few visits in the UK and one trip to the manufacturer in Boston, USA. As a result of this, we only obtained a total of 552 CT scans at a $1.56 \times 1.61 \times 5mm$ resolution. These scans were not just baggage items but also included scans to aid development such as indicated in Section 3.2. The breakdown of these is given in Table 3.1a. From these scans we obtained 1255 sub-volumes and they are summarized in Table 3.1b and Table 3.1c. It can be seen from this that there are some slight differences between the whole volume and sub-volume datasets. For example, 118 scans containing revolvers were taken but only 100 are used in the sub-volume dataset. Not all scans are used in the subsequent analysis - some were used to aid understanding of the CT environment and, once testing had commenced, were considered to be too simplistic for the task being undertaken. An example is shown in Figure 3.18 where we see a North American Arms 0.22 Mini Revolver scanned in a foam container. This revolver was not placed into any cluttered baggage environment and so it was not included in the sub-volume dataset.

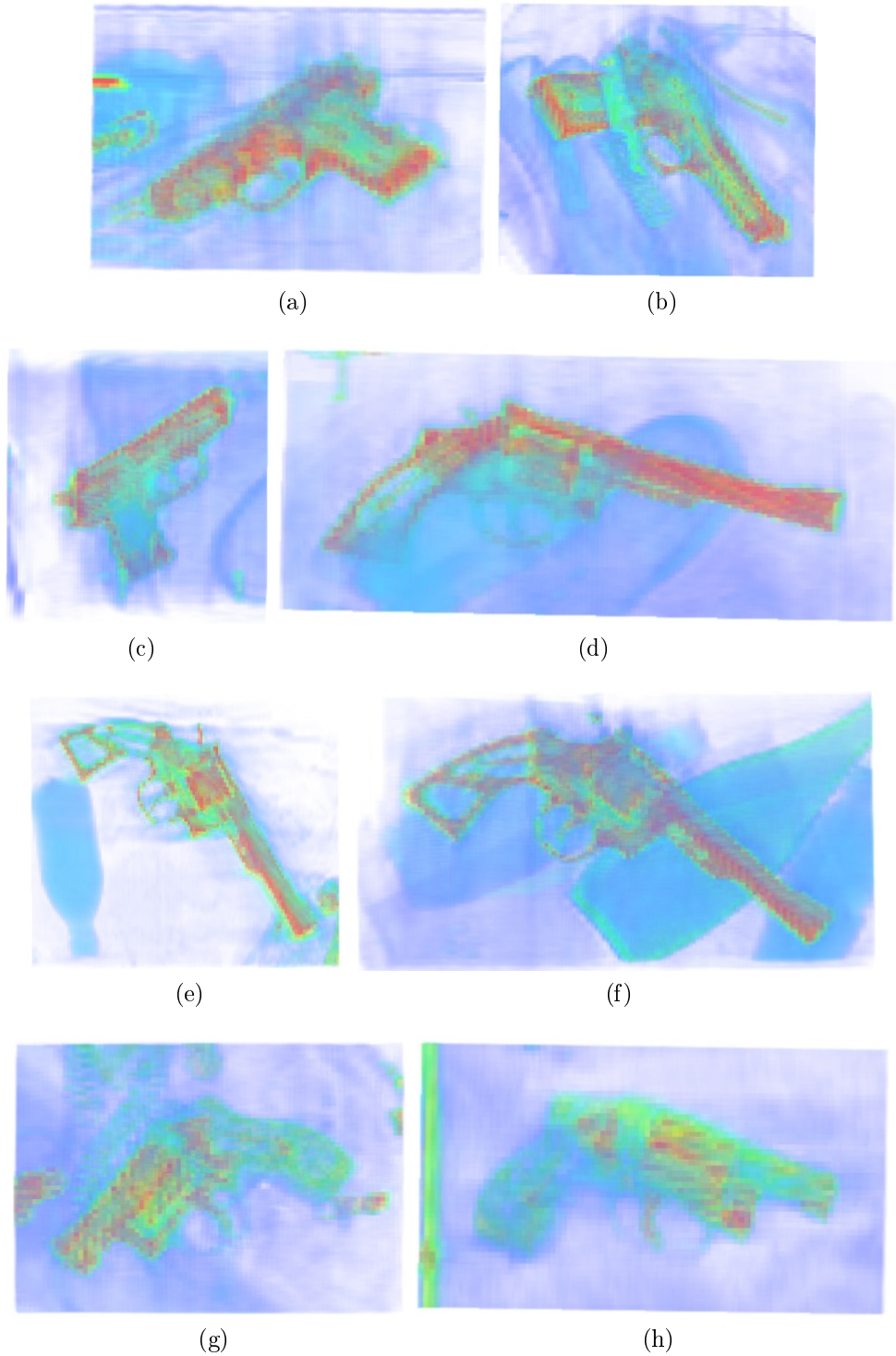


Figure 3.16: Example target item sub-volumes

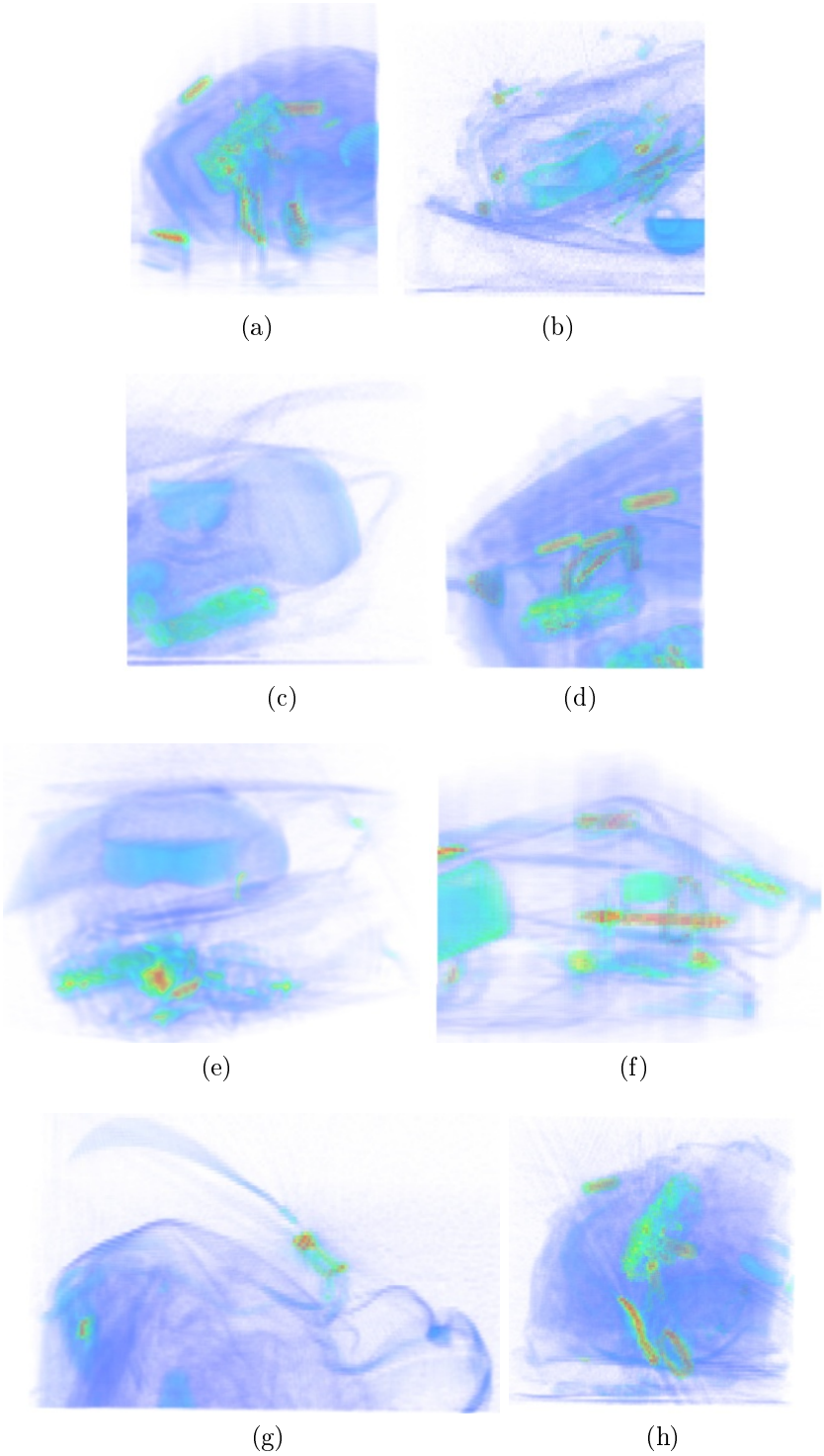


Figure 3.17: Example clutter sub-volumes

Baggage contents	Number of scans
Pistols	188
Revolvers	118
Clutter	179
Reference scans	17
Bottles	113

(a) Whole volumes

Sub-volume contents	Number of scans
Pistols	184
Revolvers	100
Clutter	971

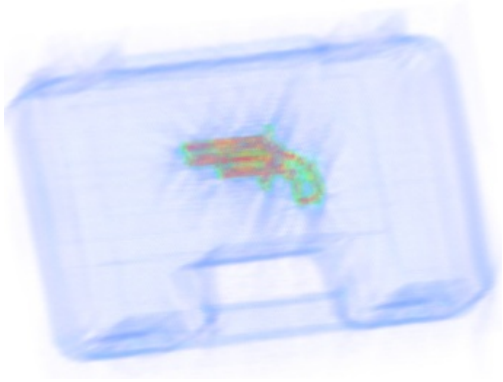
(b) Handgun sub-volumes

Sub-volume contents	Number of scans
Bottles	526
Clutter	1178

(c) Bottle sub-volumes

Table 3.1: Scan breakdowns

(a) Scanned in foam lined container



(b) Resultant CT scan

Figure 3.18: Simple pistol scan excluded from sub-volume dataset

Chapter 4

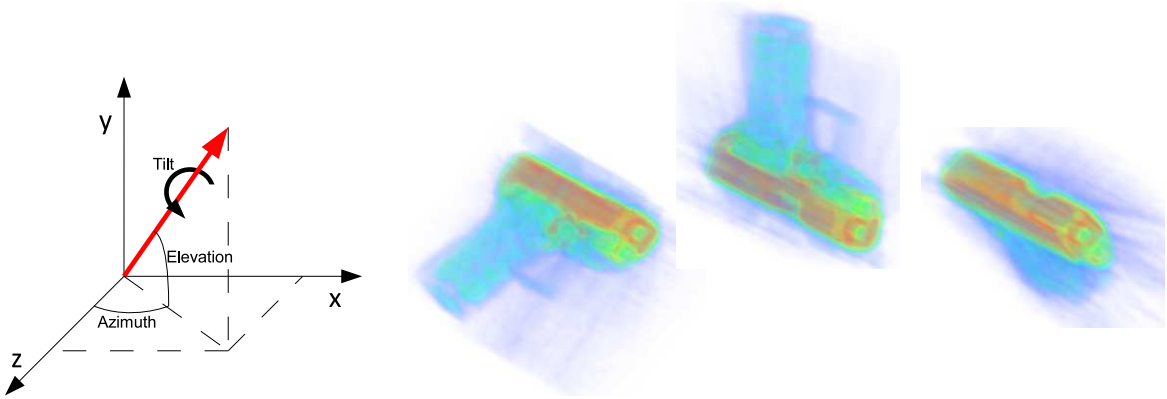
3D SIFT matching

We begin our work investigating specific-instance recognition within the obtained CT imagery. Recognition of a known item within a scene is a recognition task that will allow us to both extend the SIFT descriptor (Lowe, 2004) to three dimensions and then to explore its effectiveness in object recognition within the 3D CT-baggage imagery.

4.1 Introduction

Whilst the development of the SIFT descriptor (Lowe, 2004) has shown to be a major milestone in general object recognition in 2D imagery (Se et al., 2002; Belcher and Du, 2009; Leibe et al., 2008; Mikolajczyk and Schmid, 2005) it has yet to be successfully extended to object recognition in 3D. However, extensions for other 3D applications have been reported (Allaire et al., 2008; Cheung and Hamarneh, 2007; Ni et al., 2009; Scovanner et al., 2007). Here we extend the SIFT descriptor to 3D following the approach of Allaire et al. (2008) but with some alterations associated with the nature of our data (see Chapter 3). We examine the task of matching a known object and a volumetric scene that may contain the same object. Detection errors are evaluated to gain insight into the matching process within the CT dataset.

On important aspect of this extension is the use of *orientation* rather direction in 3D space. Figure 4.1 shows how an direction in 3D space can be defined by two angle (azimuth, elevation) whereas an orientation requires the addition of a third angle (tilt) to fully described an object state. It has been shown that the use of orientation rather than direction in 3D SIFT extensions enhances performance (Allaire et al., 2008).



(a) Definitions used in our work (b) Pistols aiming in the same direction but with differing orientation (i.e. tilt)

Figure 4.1: 3D Orientation requires three angles: azimuth, elevation and tilt

4.2 3D SIFT approach

Initially we follow the approach of Allaire et al. (2008) with additional parametric differences. Later, we extend this work to the explicit recognition of objects based on RANSAC-driven keypoint match selection, pose estimation and finally volumetric object verification.

The generation of SIFT descriptors for a volumetric image is defined in five key stages:

a) Candidate keypoint location (Section 4.2.1)

b) Rejection of poor quality locations (Section 4.2.2)

c) Refinement of keypoint location (Section 4.2.3)

d) Derivation of dominant orientation (Section 4.2.4)

e) Keypoint description (Section 4.2.5)

These steps are summarized in Figure 4.2, where we can see the progression from input volumetric image to a set of descriptors.

Throughout this discussion we use an example baggage item as shown in Figure 4.3. This shows a rucksack containing a pistol handgun, toiletries, bottles, clothing and shoes.

4.2.1 Candidate keypoint location

The first step in traditional 2D SIFT (Lowe, 2004) is the calculation of Difference of Gaussian (DoG) images. Here, given a 3D input volume $I(x, y, z)$ and a 3D Gaussian filter $G(x, y, z, \sigma^k)$ we form multi-scale Difference of Gaussian (DoG) volumes as follows:

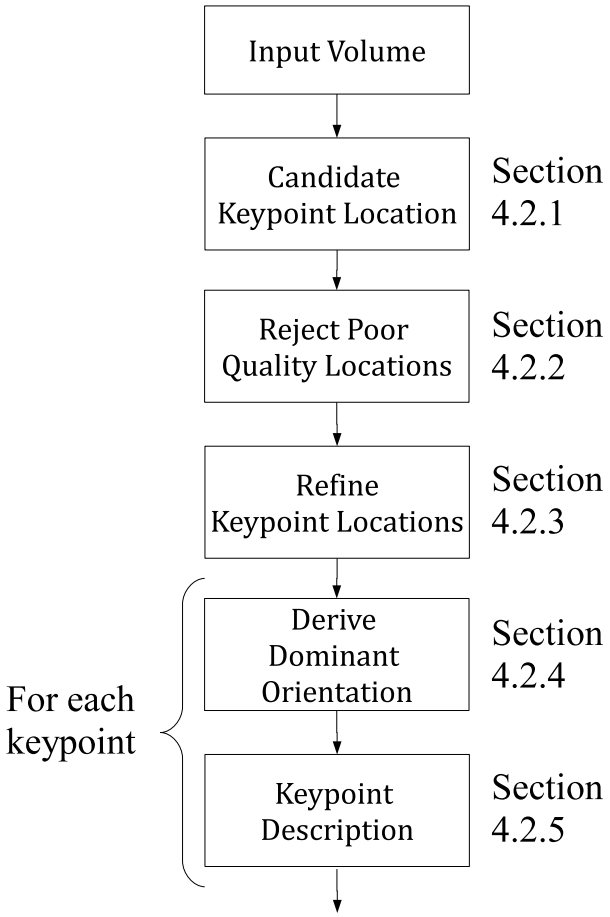


Figure 4.2: 3D SIFT algorithmic summary

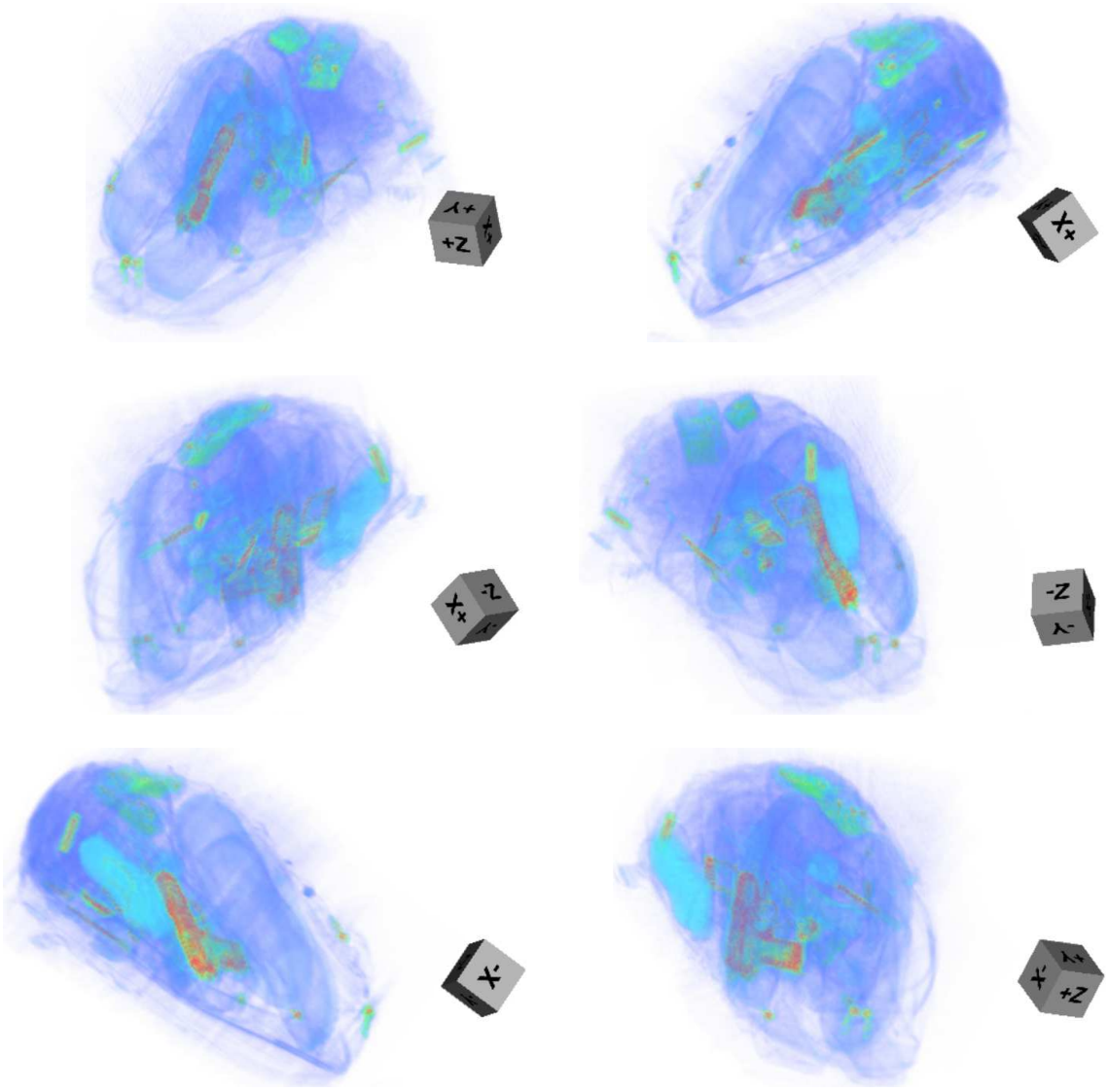


Figure 4.3: Example cluttered baggage item

$$DoG(x, y, z, k) = I(x, y, z) \otimes G(x, y, z, \sigma^k) - I(x, y, z) \otimes G(x, y, z, \sigma^{k-1}) \quad (4.1)$$

where $\otimes$ is the convolution operator, k is an integer in the range [0, 4] representing the scale index, $\sigma = \sqrt[3]{2}$ (following the work of Lowe, 2004; Allaire et al., 2008) and (x, y, z) are defined in voxel coordinates. Subsequently a three-level pyramid ($L = 0, 1, 2$) is built up by subsampling the Gaussian-filtered volume for $k = 3$ and repeating the process. Resampling with $k = 3$, coupled with the choice of σ, is such that each layer of the pyramid (L) commences with half the resolution of the previous layer ($\sigma^k = 2$ for $k = 3$). Figure 4.4 illustrates the generation of the volumetric scale-space pyramid through repeated application of Gaussian filters of different scale $\left(\sigma^k\right)$ with the resultant Difference of Gaussian scale-space pyramid. In Figure 4.5 we see the generation of Difference of Gaussian volumetric image from two Gaussian-filtered volumes for a baggage item containing a pistol, shoes and toiletries.

In a similar vein to the original 2D SIFT methodology (Lowe, 2004), DoG extrema are now located. This requires that a voxel be either a maximum or minimum when compared to its neighbouring voxels. Given that each voxel has an immediate $3 \times 3 \times 3$ local voxel neighbourhood it follows that there are 26 voxels for immediate comparison. (This is a specific case of the $N \times N \times N$ local voxel neighbourhood in which we have $N^3 - 1$ immediate neighbours). Furthermore it is also a requirement that the voxel is a maximum or minimum when compared to the 27 neighbourhood voxels in the scale space DoG volumes both above and below ($k + 1$, $k - 1$) giving a total of 80 neighbouring voxels (26 in the neighbourhood at the current scale, 27 from the scale above, 27 from the scale below). This is illustrated in Figure 4.6 where we see a local extremum voxel surrounded by 80 neighbouring voxels across both position and scale. The locations of these extrema form a candidate set of interest point locations.

4.2.2 Rejection of poor quality locations

From the candidate set of locations we first reject those that are likely to produce unstable descriptors: those for which repeatable matching will be unreliable. Additionally, in the case of CT volumes, we reject points associated with metal artefacts.

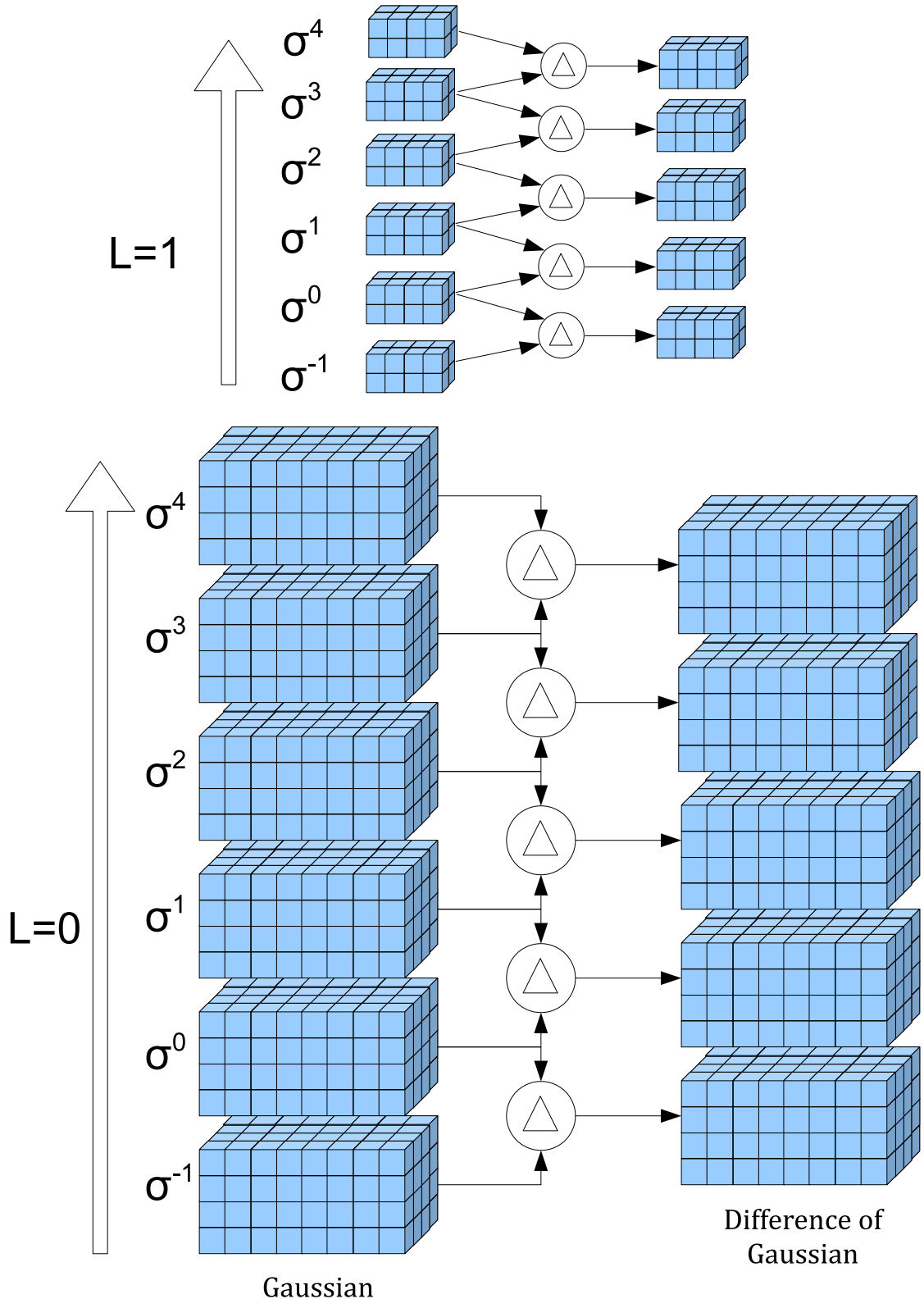


Figure 4.4: Volumetric scale-space pyramid and Difference of Gaussian generation

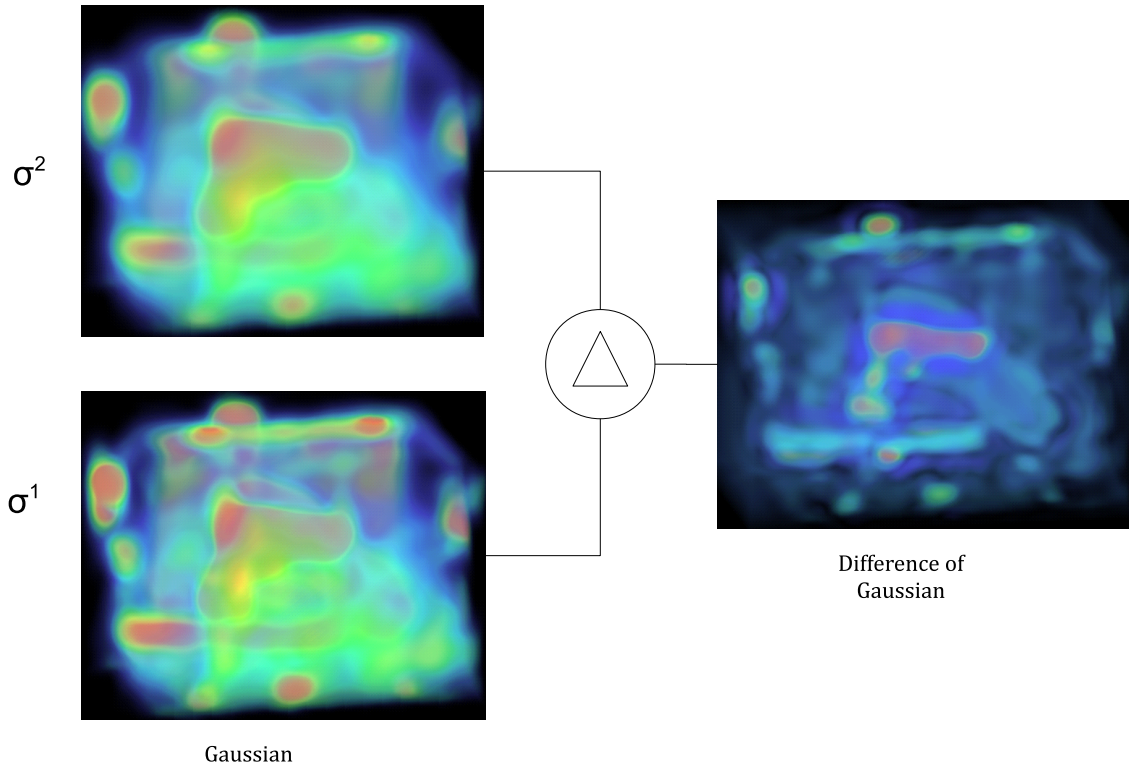


Figure 4.5: Example volumetric Difference of Gaussian generation

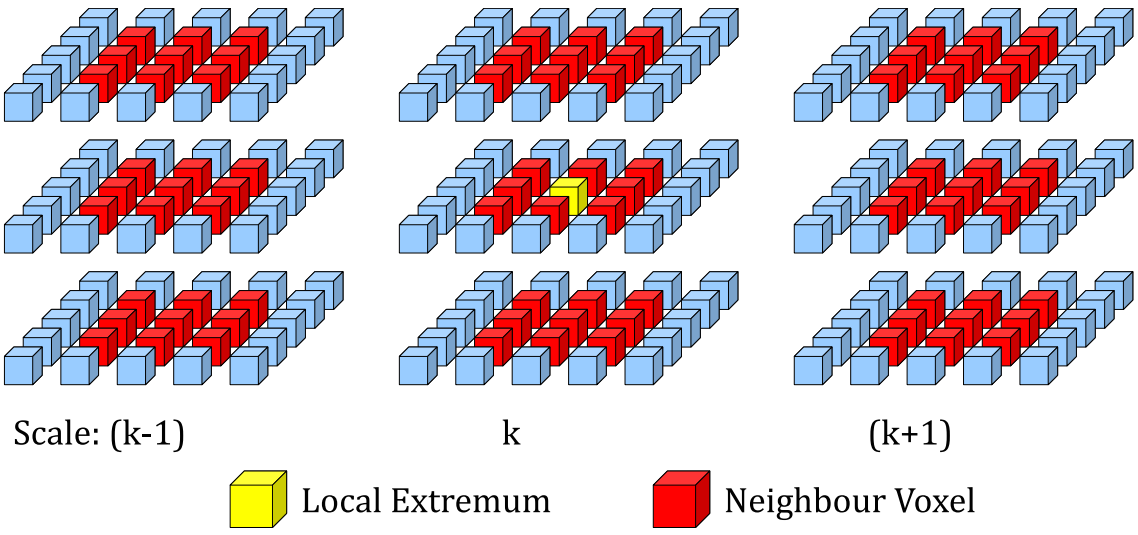


Figure 4.6: Neighbourhood voxels

4.2.2.1 Rejection: poor contrast

Points are rejected for poor contrast if their density is below a threshold, τ_c. The choice of τ_c is made through experimentation on CT volumes. We wish to detect points associated with rigid objects but we also wish to avoid the use of points associated with low density regions, as they are more easily influenced by noise and artefacts. Figures 4.7, 4.8 and 4.9 shows candidate locations (as black dots) which are retained for given values of threshold (τ_c) at each of the three pyramid levels. It can be seen in Figures 4.7a, 4.8a and 4.9a that when there is no rejection ($\tau_c = 0.00$) there are lots of interest points in noisy regions, a significant number of which are outside the baggage item arising from the streak artefacts (see Chapter 3). In Figure 4.7a we can see some structure to the locations outside the bag parallel to the z axis which may indicate that the volume interpolation in the z direction has not smoothed the volume adequately leading to a series of maxima/minima for each slice.

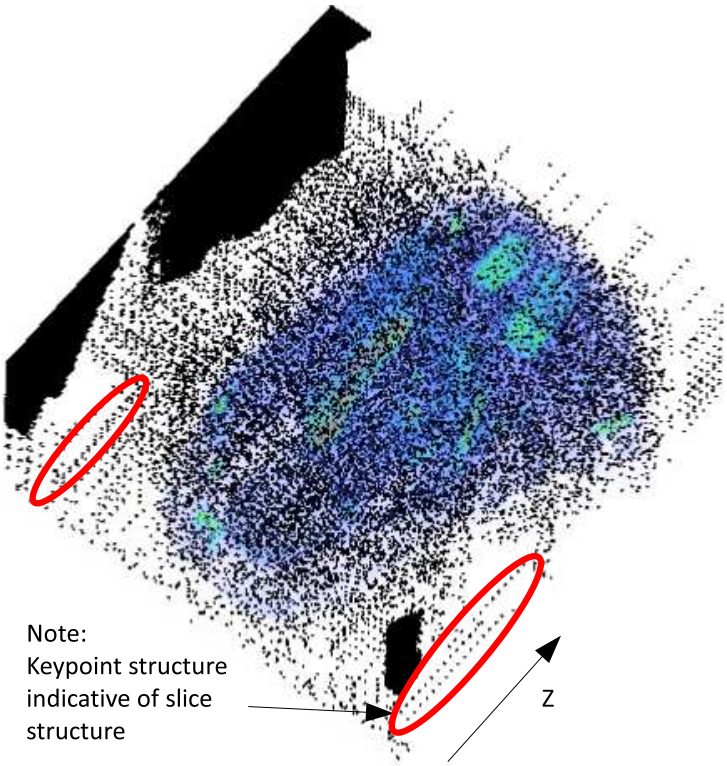
For $\tau_c = 0.05$ almost all the external interest points are removed leaving interest points generated on what appear to be significant items at all scale levels. In Figure 4.7b we see Level 0 candidate points covering most of the significant baggage contents - the shoe soles are easily seen. In Figure 4.8b we see Level 1 candidate points and in Figure 4.9b we see the Level 2 candidate points.

Increasing $\tau_c = 0.10$ reduces the number of interest points still further and we can see in Figures 4.7c, 4.8c and 4.9c that some interest points are removed from what appear to be rigid items when compared to setting $\tau_c = 0.05$ as in Figures 4.7b, 4.8b and 4.9b. This indicates that the rejection based on density is starting to remove useful interest points from significant items. This is not too surprising as the density of some of the plastic items is below 0.10.

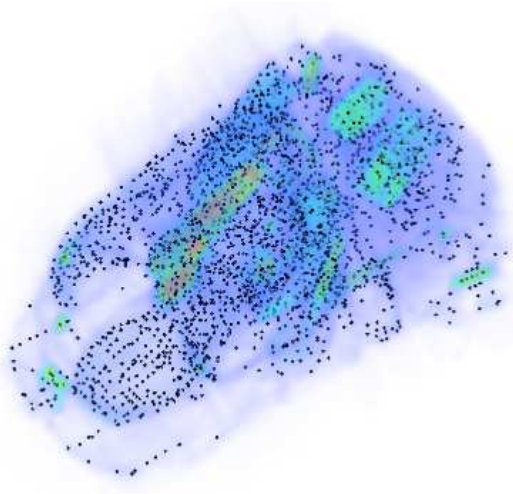
Setting $\tau_c = 0.20$ eliminates more points, leaving interest points associated with higher-density objects such as the pistol (Figure 4.9d). Experiments were performed using a variety of baggage item data that examined the quality of keypoints retained for a range of settings for τ_c. We choose to use a value of $\tau_c = 0.05$ for the rest of this work as this value can be seen to remove points associated with noise (as seen in Figures 4.7a, 4.8a and 4.9a) whilst retaining points associated with low density rigid items (e.g. shoes and belts) as shown in Figures 4.7b, 4.8b and 4.9b.

4.2.2.2 Rejection: on an edge

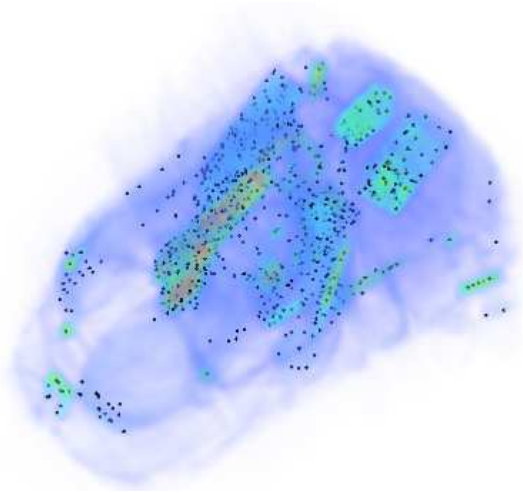
A second stage of candidate-point rejection also takes place for points which are localized on an edge: defined in 3D voxel space as a location where the principal curvatures are not similar in magnitude. Poor localization will produce lower quality



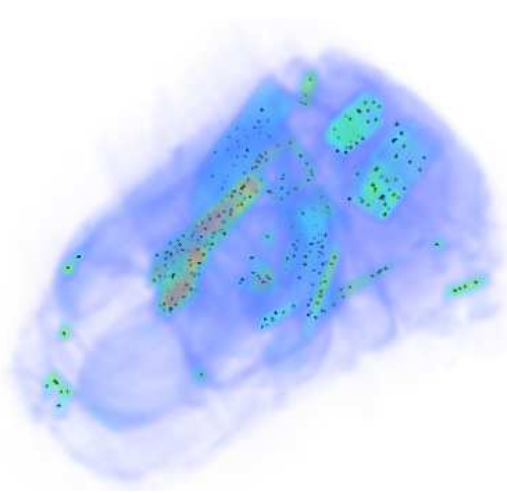
(a) $L = 0,\ \tau_c = 0.00$



(b) $L = 0,\ \tau_c = 0.05$

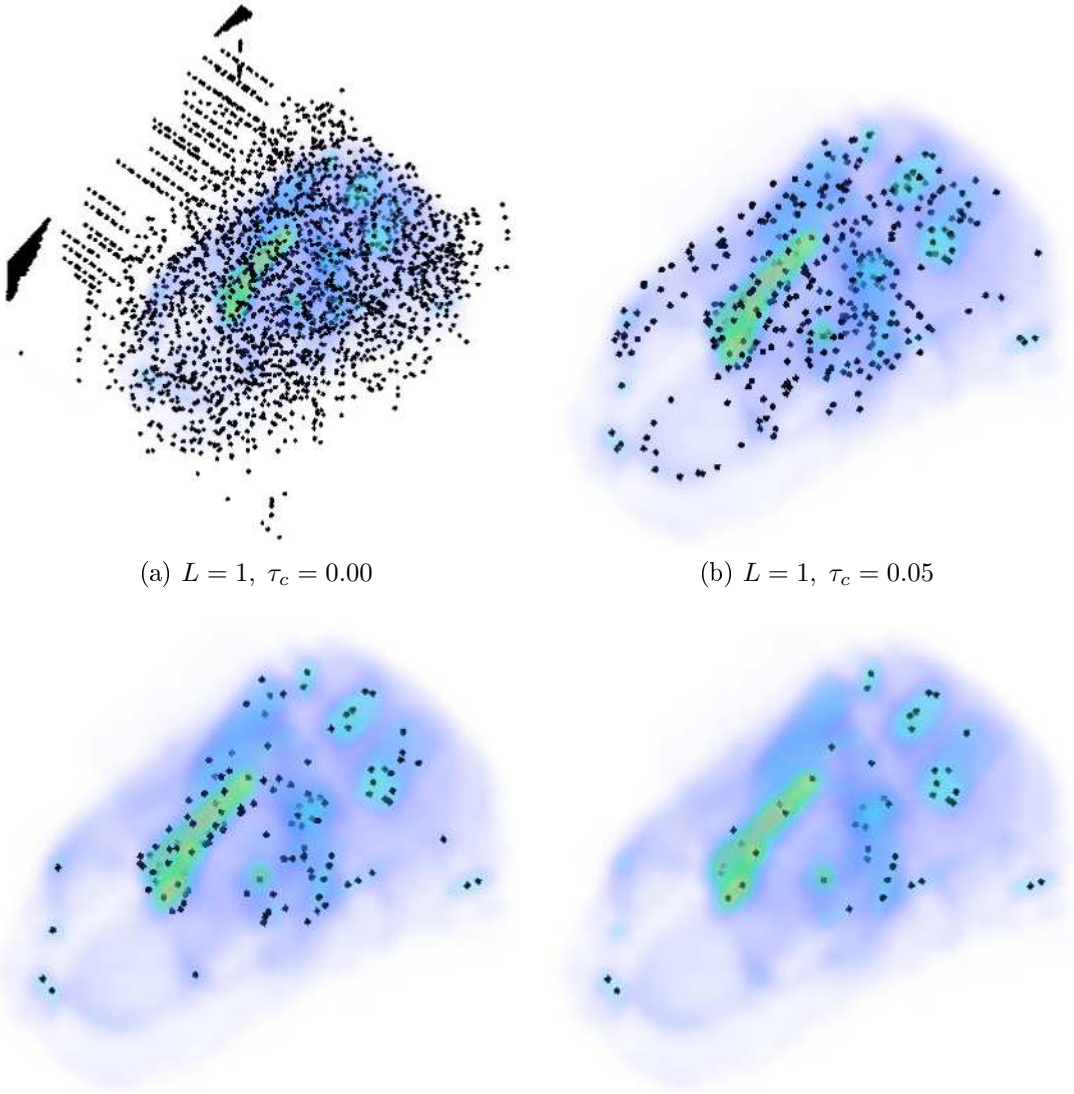


(c) $L = 0,\ \tau_c = 0.10$



(d) $L = 0,\ \tau_c = 0.20$

Figure 4.7: Level-0 Candidate-interest-point locations as τ_c is varied



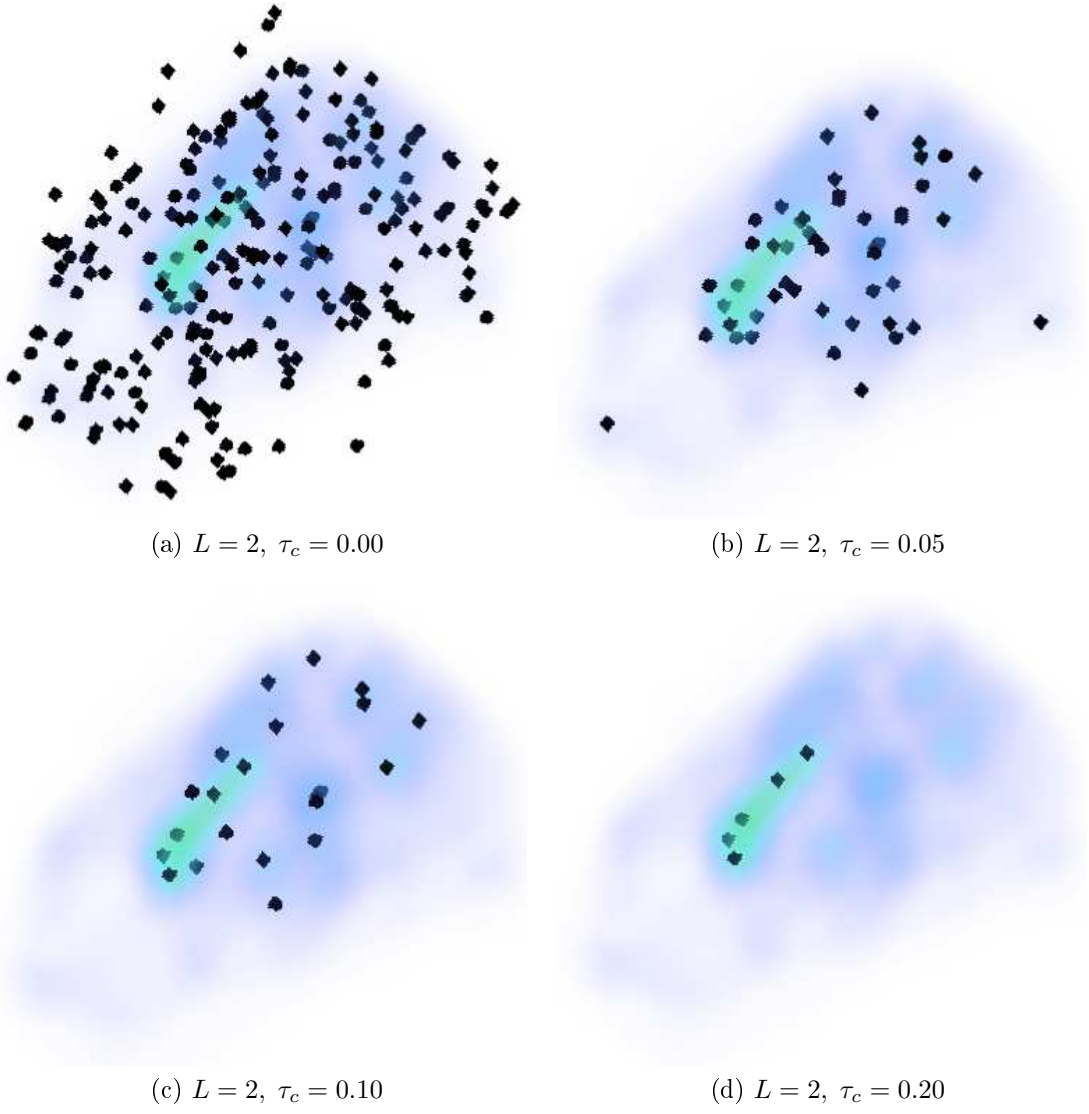
(a) $L = 1,\ \tau_c = 0.00$

(b) $L = 1,\ \tau_c = 0.05$

(c) $L = 1,\ \tau_c = 0.10$

(d) $L = 1,\ \tau_c = 0.20$

Figure 4.8: Level-1 Candidate-interest-point locations as τ_c is varied



(a) $L = 2,\ \tau_c = 0.00$

(b) $L = 2,\ \tau_c = 0.05$

(c) $L = 2,\ \tau_c = 0.10$

(d) $L = 2,\ \tau_c = 0.20$

Figure 4.9: Level-2 Candidate-interest-point locations as τ_c is varied

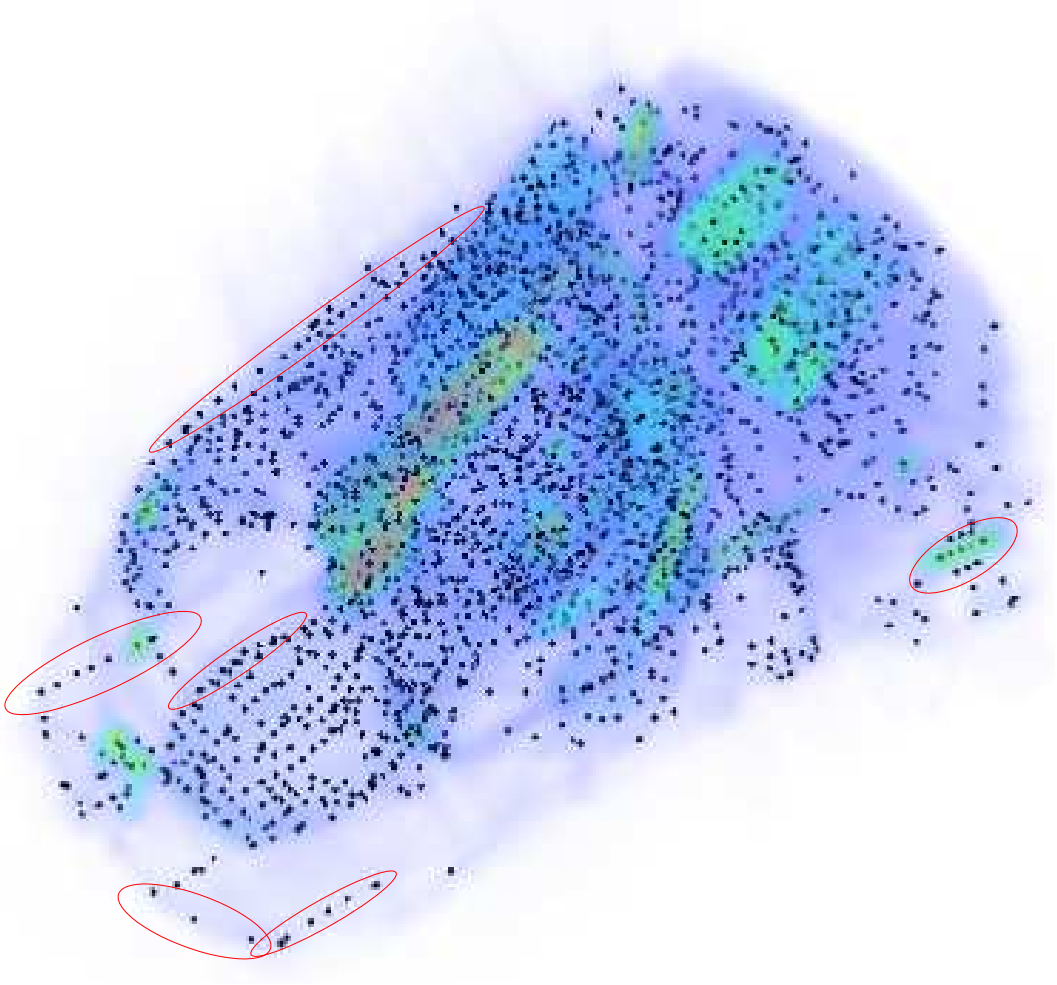


Figure 4.10: Interest points localized on edges

matches. These points are likely to produce unstable descriptors in the presence of noise. This is synonymous with the case in 2D discussed in Section 2.1.5. Figure 4.10 shows some edges that are obtained using $\tau_c = 0.05$ that we wish to remove.

Rejection is based on the local curvature in the difference of gradient volume at the candidate point, described by a 3×3 Hessian matrix:

$$H = \begin{bmatrix} D_{xx} & D_{yx} & D_{zx} \\ D_{xy} & D_{yy} & D_{zy} \\ D_{xz} & D_{yz} & D_{zz} \end{bmatrix} \tag{4.2}$$

where D_{ij} are the second derivatives in the DoG volume. The principal curvatures are proportional to the eigenvalues of H. We define the three eigenvalues of H as α, β and γ, such that

$$\alpha \geq \beta \geq \gamma. \tag{4.3}$$

Following the work of Allaire et al. (2008), we first ensure that the candidate location targets a blob-like structure. This is achieved in two stages. The first stage is to ensure that the three principal curvatures have the same sign (indicating a bright or dark blob in the DoG volume).

Both Allaire et al. (2008) and Ni et al. (2009) derive a measure to reject points using the trace and determinant of H. The trace of a matrix is the sum of the eigenvalues and the determinant is the product of the eigenvalues. This leads to:

$$Trace(H) = \alpha + \beta + \gamma = D_{xx} + D_{yy} + D_{zz} \tag{4.4}$$

$$Det(H) = \alpha\beta\gamma = D_{xx}D_{yy}D_{zz} + 2D_{xy}D_{yz}D_{xz} - D_{xx}D_{yz}^2 - D_{yy}D_{xz}^2 - D_{zz}D_{xy}^2 \tag{4.5}$$

Equation (4.4) and Equation (4.5) result in:

$$Trace(H) \times Det(H) = (\alpha + \beta + \gamma)\,\alpha\beta\gamma = \alpha^2\beta\gamma + \alpha\beta^2\gamma + \alpha\beta\gamma^2 \tag{4.6}$$

Allaire et al. (2008) also use the sum of the traces of the principal second-order minors, S_2:

$$S_2 = \beta\gamma + \gamma\alpha + \alpha\beta = D_{yy}D_{zz} - D_{yz}^2 + D_{zz}D_{xx} - D_{xz}^2 + D_{xx}D_{yy} - D_{xy}^2 \tag{4.7}$$

We can now use Equation (4.6) and Equation (4.7) to reject candidate locations where the principal curvatures are not all the same sign. If α, β and γ are the same sign (positive or negative) then S_2 will be positive as will $Trace(H) \times Det(H)$.

Given the assumption of Equation (4.3) we only need to consider two other cases:

$$\alpha > 0, \beta > 0, \gamma < 0 \tag{4.8}$$

and

$$\alpha > 0, \beta < 0, \gamma < 0 \tag{4.9}$$

We can show that $S_2 < 0$ or $Trace(H) \times Det(H) < 0$ for both these conditions. Allaire et al. (2008) state that the proof can be obtained by contradiction and we will show that now.

$$T = Trace(H) = \alpha + \beta + \gamma \tag{4.10}$$

$$\Delta = S_2 \quad = \quad \beta\gamma + \gamma\alpha + \alpha\beta \tag{4.11}$$

Rearranging Equation (4.11) gives:

$$\gamma = \frac{\Delta - \alpha\beta}{\alpha + \beta} \tag{4.12}$$

Inserting into Equation (4.10) gives:

$$T = (\alpha + \beta + \frac{\Delta - \alpha\beta}{\alpha + \beta})\, \alpha\beta \left(\frac{\Delta - \alpha\beta}{\alpha + \beta} \right) \tag{4.13}$$

In order to prove by contradiction we first assert that $T > 0$ and $\Delta > 0$ under both conditions (4.8) and (4.9).

Hence:

$$(\alpha + \beta + \frac{\Delta - \alpha\beta}{\alpha + \beta})\, \alpha\beta \left(\frac{\Delta - \alpha\beta}{\alpha + \beta} \right) > 0$$

$$\frac{1}{\alpha + \beta}((\alpha + \beta)^2 + \Delta - \alpha\beta)\, \alpha\beta \left(\frac{\Delta - \alpha\beta}{\alpha + \beta} \right) > 0 \tag{4.14}$$

For condition (4.8) it follows that for (4.12) we must have $\Delta - \alpha\beta < 0$

So for (4.14) to hold true

$$(\alpha + \beta)^2 + \Delta - \alpha\beta < 0 \tag{4.15}$$

$$\alpha^2 + \alpha\beta + \beta^2 + \Delta < 0$$

Which is only possible if $\triangle < 0$, which contradicts our earlier assertion.

For the condition of Equation (4.9) it follows that for Equation (4.12) we consider the following two states:

1. if $|\alpha| > |\beta|$ then $\alpha + \beta > 0$ so $\Delta - \alpha\beta < 0$

$\Delta < \alpha\beta$ but $\alpha\beta < 0$ in this case which implies that $\Delta < 0$ which again contradicts the earlier assertion.

2. if $|\alpha| < |\beta|$ then $\alpha + \beta < 0$ so $\Delta - \alpha\beta > 0$ from 4.12 as $\gamma < 0$.

Using these in Equation (4.14) again gives Equation (4.15) which is only possible if $\triangle < 0$, which again contradicts our earlier assertion.

Given the assertion that $T > 0$ and $\Delta > 0$ has now been shown to be false when α, β and γ do not have the same sign we can use the two following rejection criteria:

$$\text{Reject point if } S_2 \leq 0 \tag{4.16}$$

$$\text{Reject point if } Trace(H) \times Det(H) \leq 0 \tag{4.17}$$

Figure 4.11 shows the impact of these two rejection criteria. Figures 4.11a, 4.11b and 4.11c show the interest points following the initial selection stage (Section 4.2.1) where we can see significant number of interest points along edges (most noticeable in Figure 4.11a and 4.11b along the shoe outlines). Figures 4.11d, 4.11e and 4.11f show the interest points remaining in the same volume following the initial rejection for non-blob-like points (see Section 2.1.2). There has been a dramatic reduction in the number of points such that most of the edge-located points have been removed.

We are now left with blobs, edges and ridges (Section 2.1.5.1). The second stage of rejection targets edges and ridges as they yield poorly localized points whose descriptors give poor spatial matching.

As stated in both (Allaire et al., 2008; Ni et al., 2009) we can express the eigenvalue relationships as follows:

$$\beta = s\alpha \tag{4.18}$$

$$\gamma = r\alpha \tag{4.19}$$

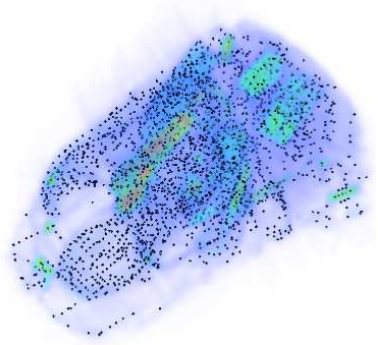
$$Trace(H) = \alpha + \beta + \gamma = \alpha + s\alpha + r\alpha = \alpha(1 + s + r) \tag{4.20}$$

$$Det(H) = \alpha\beta\gamma = \alpha s\alpha r\alpha = \alpha^3 sr \tag{4.21}$$

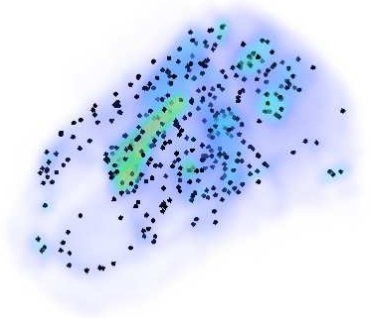
Then

$$\frac{Trace(H)^3}{Det(H)} = \frac{[\alpha\,(1 + s + r)]^3}{\alpha^3 sr} = \frac{(1 + s + r)^3}{sr} \tag{4.22}$$

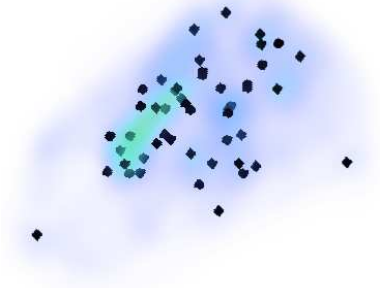
Given that s is smaller than r, we can rewrite 4.22 as an inequality by substitution in Equation (4.22). We first observe:



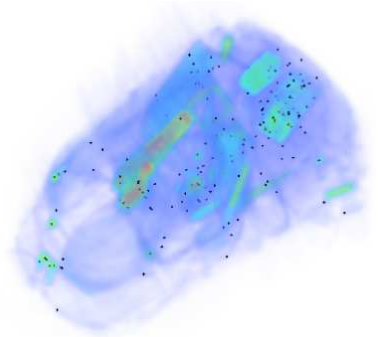
(a) Level-0, $\tau_c = 0.05$, No edge rejection



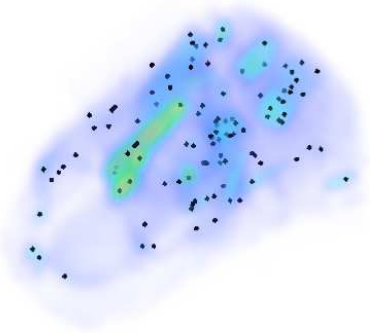
(b) Level-1, $\tau_c = 0.05$, No edge rejection



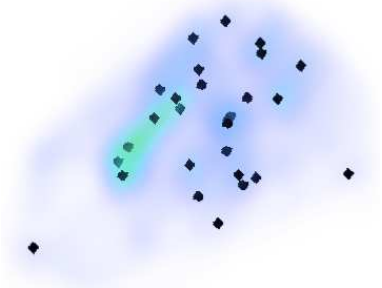
(c) Level-2, $\tau_c = 0.05$, No edge rejection



(d) Level-0, First stage rejection



(e) Level-1, First stage rejection



(f) Level-2, First stage rejection

Figure 4.11: Non-blob rejection

$$s < r \;\Rightarrow\; \frac{(1+s+r)^3}{sr} < \frac{(1+2r)^3}{r^2} \tag{4.23}$$

so that:

$$\frac{Trace(H)^3}{Det(H)} < \frac{(1+2r)^3}{r^2} \tag{4.24}$$

If we wish to restrict our selection to cases where the eigenvalue magnitudes $(|\alpha|, |\beta|, |\gamma|)$ are in a known range then we can set a threshold r_{max} for r which this leads us to:

$$\frac{Trace(H)^3}{Det(H)} < \frac{(1+2r_{max})^3}{r_{max}^2} \tag{4.25}$$

The rejection based on the similarity of eigenvalue magnitude is then:

$$\text{Reject point if } \frac{Trace^3(H)}{Det(H)} > \frac{(2\tau_e+1)^3}{(\tau_e)^2} \tag{4.26}$$

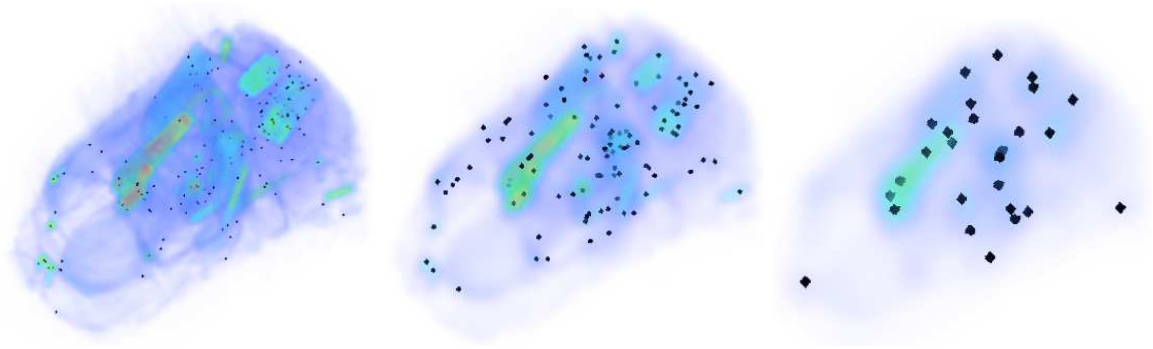
where the parameter τ_e defines how similar the eigenvalues should be.

4.2.2.3 Rejection: selection of τ_e value

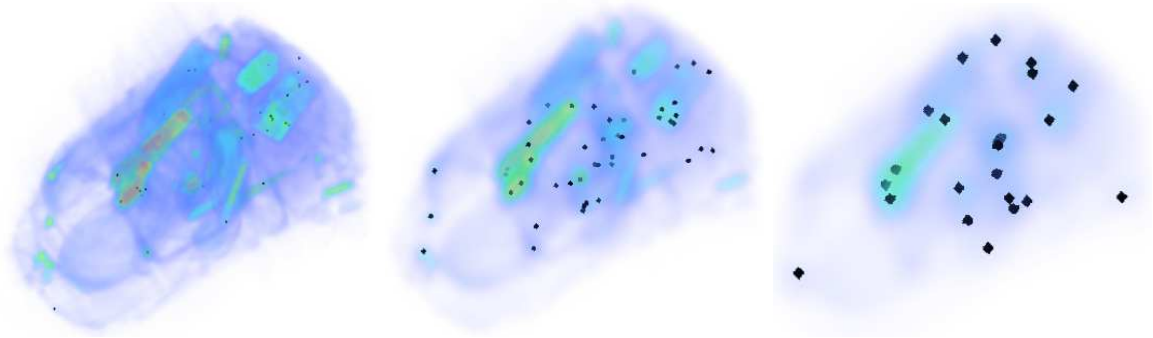
Choosing a value for τ_e that maximizes the rejection of poor quality points whilst retaining as many good quality points is the next step.

Figure 4.12 shows the interest points at each pyramid level that remain for a fixed $\tau_c = 0.05$ as τ_e is varied. In Figures 4.12a, 4.12b and 4.12c we see the interest points that remain following the rejection based on low local density value ($\tau_c = 0.05$) followed by rejection if $S_2 < 0$ (see Equation (4.16)). Figures 4.12d, 4.12e and 4.12f show the introduction of rejection using Equation (4.26) with a setting of $\tau_e = 5.0$ where we can see significantly fewer interest points than in Figures 4.12a, 4.12b and 4.12c. Figures 4.12g, 4.12h and 4.12i show the interest points that remain for a setting of $\tau_e = 40.0$ which shows more points remaining than for $\tau_e = 5.0$.

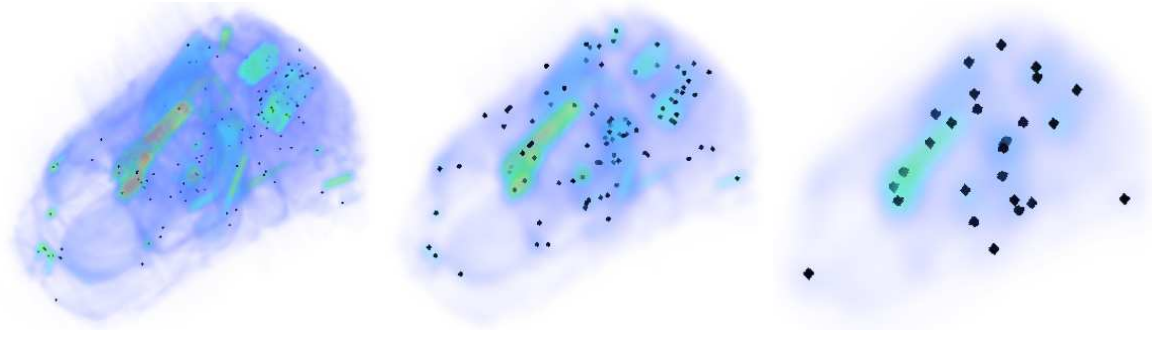
Allaire et al. (2008) used a value of $\tau_e = 5.0$ for their CT data and a value of $\tau_e = 20.0$ for both magnetic resonance and cone-beam CT imagery. Ni et al. (2009) used a value of $\tau_e = 15.0$ for their ultrasound imagery. We use a less discriminatory value of $\tau_e = 40.0$ as we wish to ensure that we keep salient feature points at the expense of an increased level of noise. Setting $\tau_e = 40.0$ in Equation (4.26) implies rejection of locations where $\frac{Trace^3(H)}{Det(H)} > 332.15$.



(a) Level 0, first stage rejection (b) Level 1, first stage rejection (c) Level 2, first stage rejection



(d) Level 1, $\tau_e = 5.0$ (e) Level 2, $\tau_e = 5.0$ (f) Level 3, $\tau_e = 5.0$



(g) Level 1, $\tau_e = 40.0$ (h) Level 2, $\tau_e = 40.0$ (i) Level 3, $\tau_e = 40.0$

Figure 4.12: Varying τ_e

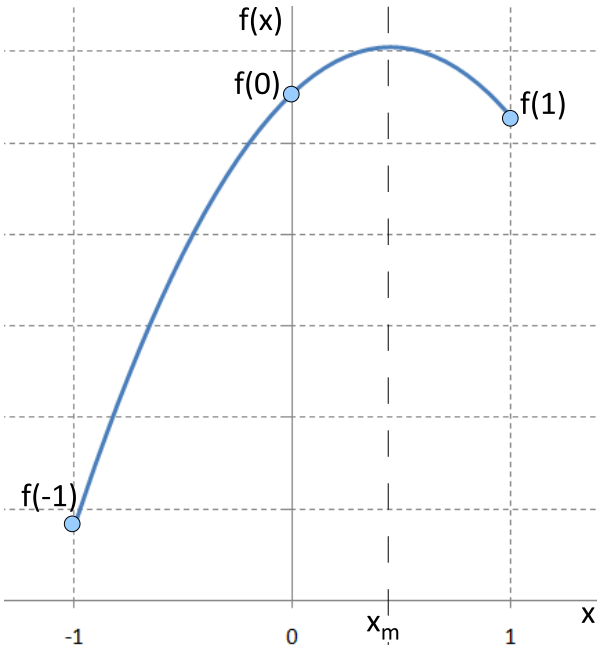


Figure 4.13: Parabolic curve fitting to estimate location of maxima/minima

4.2.3 Location refinement

The candidate points are defined by integer voxel locations. We wish to obtain a sub-voxel refinement on this in line with Lowe (2004) and Allaire et al. (2008).

Brown and Lowe (2002) extended the original SIFT algorithm from using integer pixel locations (Lowe, 1999) to sub-pixel locations with a considerable improvement in detection. We implement a sub-voxel location refinement by fitting a parabola along each dimension (xyz) and calculating the max/min accordingly, as shown in Figure 4.13. Three equidistant points are chosen around the local maxima/minima and a parabola fitted. The location of the maxima/minima of the parabola is calculated and used as the refined location for the interest point along the respective axis. This method differs from the work of Brown and Lowe (2002) and Lowe (2004) and is an area for future work.

From Figure 4.13 the maxima/minima location is estimated as:

$$x_m = \frac{f(-1) - f(1)}{2(f(-1) + 2f(0) + f(1))} \tag{4.27}$$

Relating this to the position in space and scale, for an initial location estimate of $(x_c, y_c, z_c, \sigma_c)$ with voxel value $I(x_c, y_c, z_c, \sigma_c)$ we obtain a sub-voxel maxima/minima location estimate $(\hat{x}, \hat{y}, \hat{z}, \hat{\sigma})$ through repeated application of Equation (4.27) along each direction:

$$\hat{x} = \frac{I(x_c - 1, y_c, z_c, \sigma_c) - I(x_c + 1, y_c, z_c, \sigma_c)}{2(I(x_c - 1, y_c, z_c, \sigma_c) + 2I(x_c, y_c, z_c, \sigma_c) + I(x_c + 1, y_c, z_c, \sigma_c))} \tag{4.28}$$

$$\hat{y} = \frac{I(x_c, y_c - 1, z_c, \sigma_c) - I(x_c, y_c + 1, z_c, \sigma_c)}{2(I(x_c, y_c - 1, z_c, \sigma_c) + 2I(x_c, y_c, z_c, \sigma_c) + I(x_c, y_c + 1, z_c, \sigma_c))} \tag{4.29}$$

$$\hat{z} = \frac{I(x_c, y_c, z_c - 1, \sigma_c) - I(x_c, y_c, z_c + 1, \sigma_c)}{2(I(x_c, y_c, z_c - 1, \sigma_c) + 2I(x_c, y_c, z_c, \sigma_c) + I(x_c, y_c, z_c + 1, \sigma_c))} \tag{4.30}$$

$$\hat{\sigma} = \frac{I(x_c, y_c, z_c, \sigma_c - 1) - I(x_c, y_c, z_c, \sigma_c + 1)}{2(I(x_c, y_c, z_c, \sigma_c - 1) + 2I(x_c, y_c, z_c, \sigma_c) + I(x_c, y_c, z_c, \sigma_c + 1))} \tag{4.31}$$

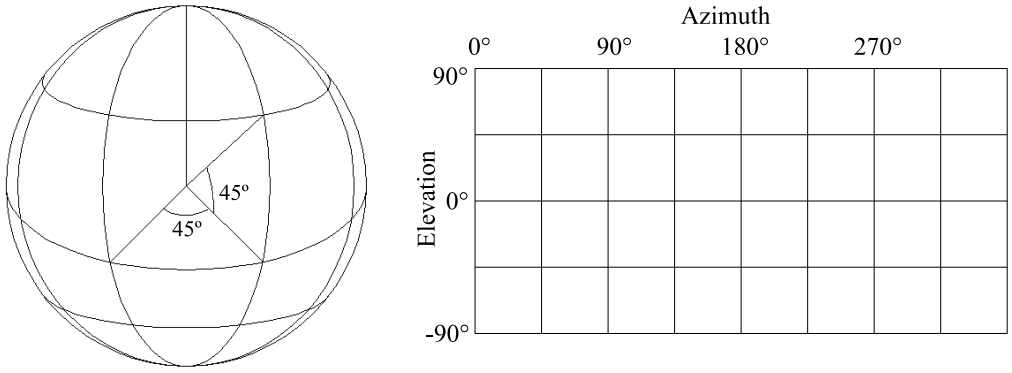
In practice we choose not to refine the scale position ($\hat{\sigma}$) - keeping the scale to the nearest in the scale pyramid that results in the maxima/minima. Given that the CT imagery expresses voxels as real-world measurements the necessity to cope with large scale transformations in the apparent size of objects of interest is not applicable.

4.2.4 Keypoint orientation

Once a keypoint location is determined, the volume gradients are examined in a two-stage process to locally establish an invariant orientation in the subsequent description. A *direction* in 3D space is defined by the azimuth and elevation angles whereas an *orientation* is defined by the addition of a third angle around the direction axis: tilt (see Figure 4.1).

The first step is to determine the dominant *direction* for the keypoint. We resample the Gaussian-filtered volumes according to ($\hat{x}, \hat{y}, \hat{z}$,$\hat{\sigma}$) to relocate the keypoint to an integer location before calculating the volume gradients.

We then form a 2D histogram by grouping the volume gradients in 32 bins, following Allaire et al. (2008), which divide azimuth and elevation into 45° sections, as shown in Figure 4.14a (sphere) and Figure 4.14b (resulting 2D histogram bins). A regional weighting is applied to the gradients according to their voxel distance from the keypoint location: we apply a Gaussian weighting of $exp\left[-(2r/R_{max})^2\right]$ for voxels a distance r from the keypoint location. Points further than R_{max} voxels from the location are ignored in the current formulation. This is done to restrict the



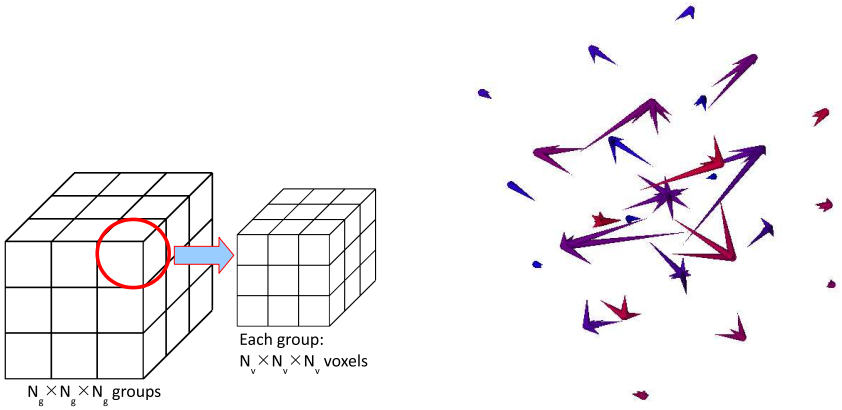
(a) Splitting the azimuth and elevation into 45° bins

(b) Resultant 2D histogram plane

Figure 4.14: Direction histogram

contribution of voxels to those in the local region. From a geodetic viewpoint (Figure 4.14a) it can be seen that histogram bins near the *equator* in this formulation are larger than those at the poles and this biases the resulting histogram to emphasize the equatorial bins. This bias is compensated for by normalizing each histogram bin by its solid angle (Scovanner et al., 2007). The output histogram is then smoothed using a Gaussian filter to limit the effects of noise and the dominant directions are determined by searching for peaks and are refined using interpolation in a similar approach to Section 4.2.3. Peaks in this 2D histogram within 80% of the largest peak are empirically retained as possible secondary directions in line with the findings of Lowe (2004).

The second step is to determine the *orientation* by calculating the tilt angle for each derived direction. This is achieved by re-orientating the volume around the keypoint using the obtained values for azimuth and elevation such that the dominant direction is aligned along the x-axis. A 1D histogram that resolves the gradients orthogonal to the dominant direction (i.e. in the yz plane) is then calculated. This histogram is again built in 45° bins using the same regional-weighting method as for the direction histogram. Peaks in the tilt histogram are used, with interpolation, to derive an estimate of keypoint tilt. Again, peaks within 80% of the largest peak are retained to give secondary orientations. Overall, in this formulation, we see that keypoints may have more than one possible orientation that will require description.



(a) Voxel groupings for descriptor (b) $3 \times 3 \times 3$ 3D SIFT descriptor

Figure 4.15: 3D SIFT descriptor formulation

4.2.5 Keypoint description

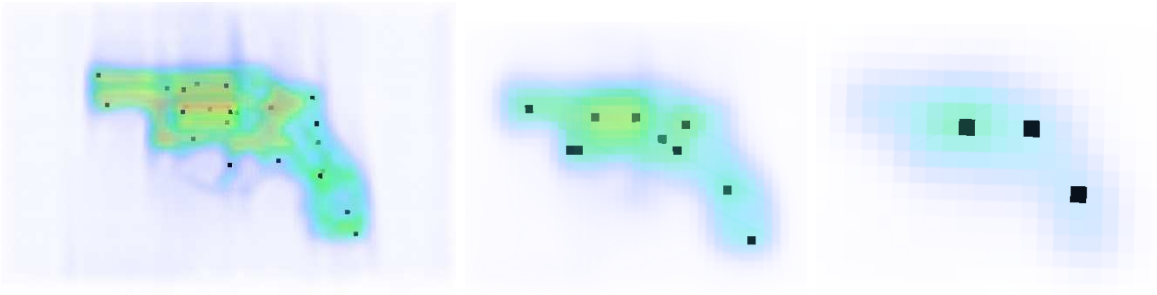
Once the orientation has been determined, the point of interest can be described. In our case we build a $N_g \times N_g \times N_g$ grid of gradient histograms ($N_g = 3$), with each histogram being computed from a $N_v \times N_v \times N_v$ voxel grouping ($N_v = 3$) as shown in Figure 4.15a. Each gradient histogram is derived by splitting both azimuth and elevation into 45° bins, as described in Section 4.2.4. Consequently, each descriptor, normalized to unity, contains $N_g^3 \times 8 \times 4$ elements. The final visualization of such a descriptor is shown in Figure 4.15b as a 3D grid of gradient histograms. This follows the 2D approach of Lowe (2004) and the 3D extension of Allaire et al. (2008).

4.3 Object identification

Following from our extension of SIFT into a 3D voxel formulation, we follow a traditional route of object identification (Lowe, 2004) where we search for a reference object in a scene and use a RANSAC-based formulation to identify a given set of consistent matches (Szeliski, 2010).

A separate scan of the item of interest being considered was taken, from which the item is cropped to provide a reference volume. This reference volume is then subjected to the 3D SIFT generation process creating a reference descriptor set. Figure 4.16 shows this reference volume with the location of its keypoints at the 3 resolutions in the earlier scale-space pyramid (see Section 4.2.1/Figure 4.4). It should be noted that this reference is also subject to the CT artefacts and resolution issues previously discussed (Chapter 3).

Here, each example baggage item, when processed as described, will produce a corresponding set of candidate descriptors. The reference descriptors are compared



(a) $L = 0$, $2.5mm$ voxels (b) $L = 1$, $5mm$ voxels (c) $L = 2$, $10mm$ voxels

Figure 4.16: Revolver reference item keypoints (in black) at different scale-space pyramid resolutions

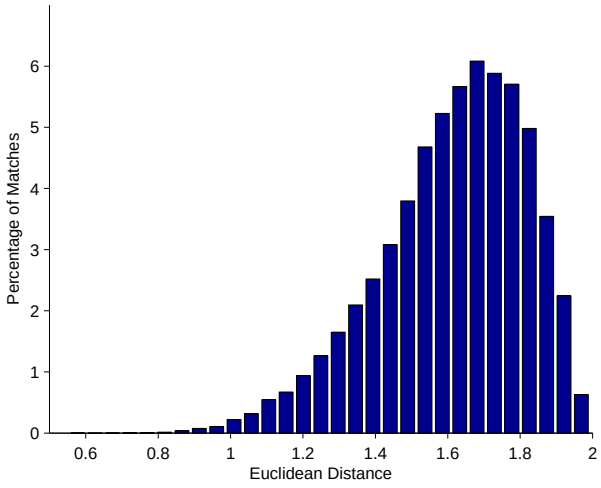
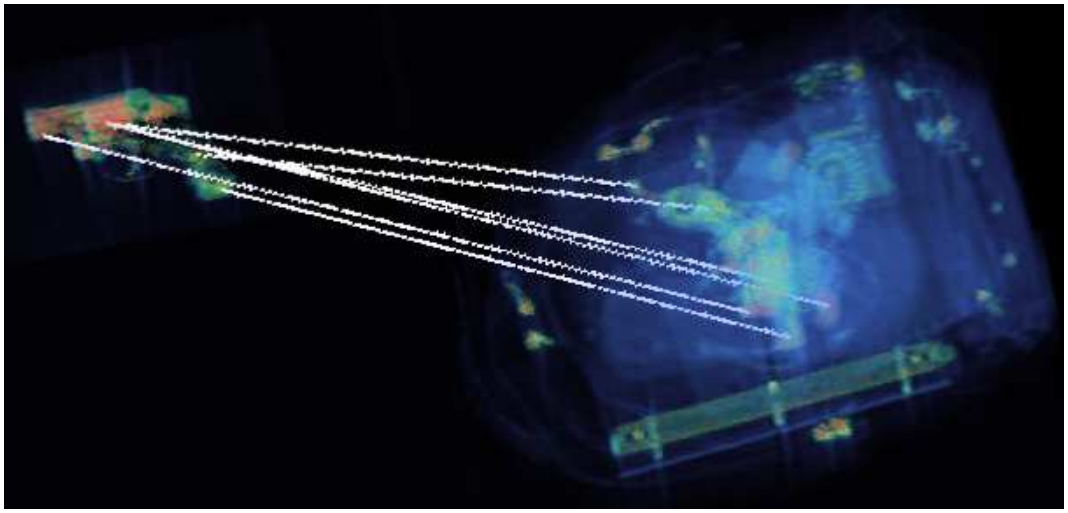


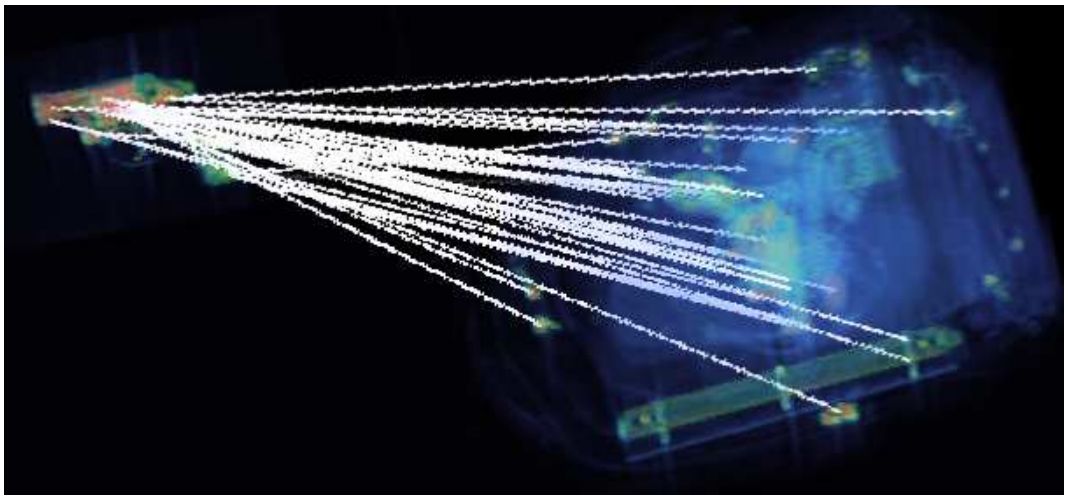
Figure 4.17: Histogram of Euclidean distances between reference-object descriptors and candidate-bag descriptors for the revolver in Figure 4.16 and baggage item in Figure 4.18.

to the candidate descriptors by recording the Euclidean distance between them (Lowe, 2004). Figure 4.17 shows a histogram of the Euclidean distances measured in a typical candidate bag. A hard rejection criterion is employed on these distance values using a fixed threshold, τ_m, such that only candidate/reference pairs with distances below τ_m are retained as an array of possible 3D SIFT matches. Figure 4.18 shows matches from a reference object to a candidate bag as the decision threshold, τ_m, is varied and it can be seen that the number of matches (both true and false) increase as τ_m increases.

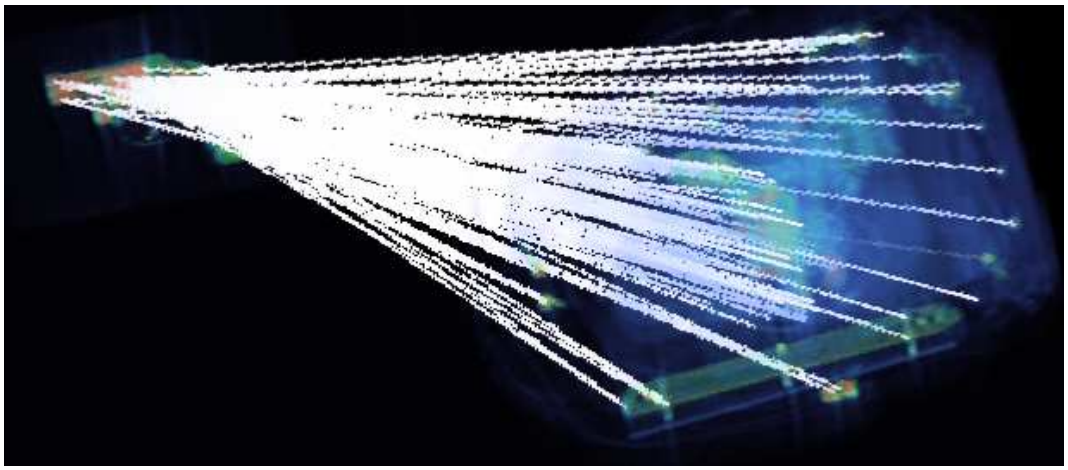
Given the large number of possible false matches in this formulation (Figure 4.18) we make use of RANSAC (Fischler and Bolles, 1981) to find an optimal match between the reference item descriptors and a subset of the candidate descriptors. The RANSAC methodology uses random selection of data from the candidate set followed by a verification stage to derive a transform of the reference object into the



(a) $\tau_m = 0.8$



(b) $\tau_m = 0.9$



(c) $\tau_m = 1.0$

Figure 4.18: Candidate matches between reference object and candidate bag for different settings of τ_m

candidate baggage space that achieves a 'best fit' between the reference item and the candidate baggage. RANSAC has been shown to cope well in the presence of significant outliers (here highly prevalent due to noise). This RANSAC formulation is used to select a set of three possible matches from which a 3D transformation is derived using singular value decomposition (Arun et al., 1987). This calculates the transform required to back-project the reference keypoints into the baggage item. An additional constraint is used to enforce consistency between the relative distances of the transformed reference set and the selected candidate match points: any relative distance errors greater than δ_r ($\delta_r = 10mm$) will result in the transformation being rejected. The value of δ_r was chosen through empirical evaluation of the quality of matches chosen by the RANSAC approach.

If this relative distance criterion is passed, a secondary verification is performed using a comparison of CT reference to candidate object density. All locations within the reference object with density above a threshold τ_d are compared using L_1 distance on a voxel by voxel basis.

If there are N_m verification voxels derived from the reference object then the match measure is defined by:

$$M = \frac{\sum_{m=1}^{N_m} |\,(I_m - I_c)\,|}{\sum_{m=1}^{N_m} I_m} \tag{4.32}$$

where I_m is the density at the m^{th} verification voxel and I_c is the density of the closest voxel in the baggage item given by the transformation of the m^{th} verification point into the baggage item.

This is recorded as the verification match metric. Combined with RANSAC this is used to identify the best candidate match within a complex volume for a given reference item.

4.4 Results

Results based on our approach are presented with a set of volumes created using the process outlined in Chapter 3. For the SIFT descriptor we empirically use: $N_g = 3$ and $N_v = 3$ (Section 4.2.5) with $R_{max} = 9$ for the Gaussian weighting (Section 4.2.4). A setting of $\tau_m = 1.2$ was chosen for the matching decision threshold (Section 4.3) as a value that maintained enough correct matches in the candidate sets whilst rejecting enough points to make the RANSAC selection find a correct match in a timely manner.

A number of target items were used to evaluate the target recognition in a variety

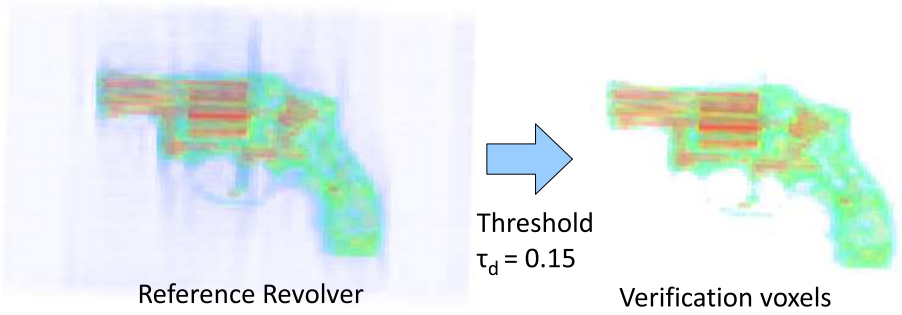
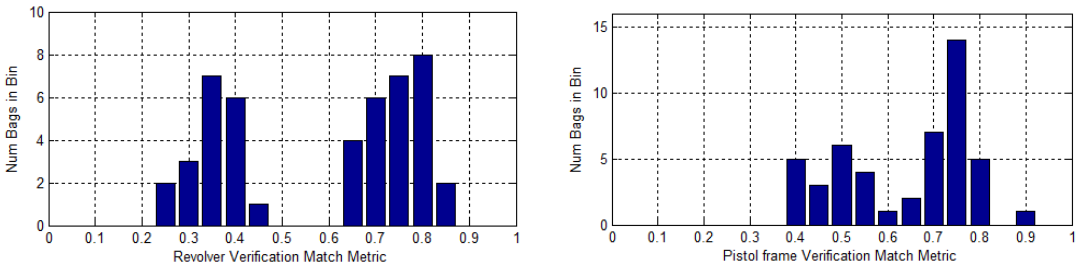


Figure 4.19: Revolver verification voxels

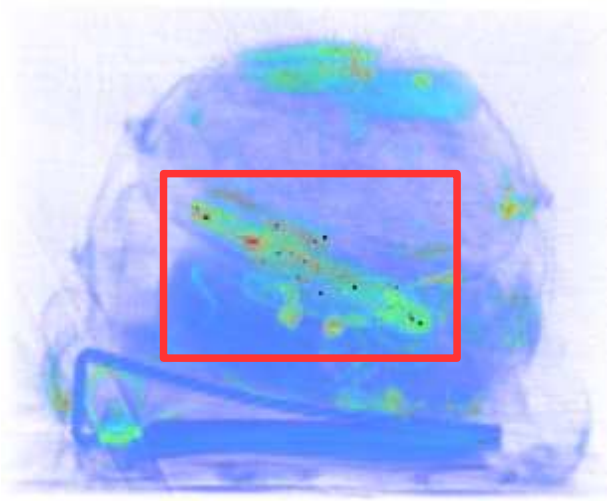


(a) Revolver match metric (b) Pistol-frame match metric

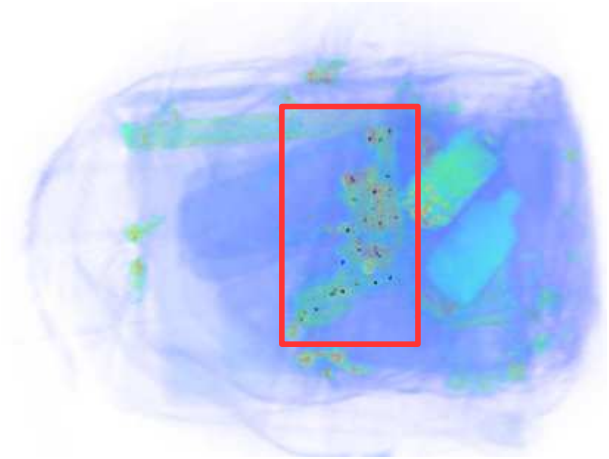
Figure 4.20: Histogram of target verification match metric results

of cluttered baggage CT images (see Chapter 3 for an overview). Firstly a revolver-type handgun (.357 Magnum, Figure 4.18/ Figure 4.16) was concealed in various baggage items producing a set of 21 3D CT scan images. An additional 25 bag set of negative (target not present) scans were also generated. Over this combined set (46 CT baggage scans) the match metric (Section 4.3) was evaluated for each bag using a reference object density threshold of $\tau_d = 0.15$. Figure 4.19 shows the verification volume for the revolver target where we can see that a significant number of the artefacts outside the object have been removed leaving voxels with density greater than 0.15. In Figure 4.20a we see a histogram of the match metric result over this set which shows two distinct regions (i.e. peaks) from which a decision threshold on this distribution can be set to determine target identification. Using a match metric threshold τ_i ($\tau_i = 0.55$) over this distribution (Figure 4.20a) yields the target-detection result shown in Table 4.1a. Here (Table 4.1a) we see a strong result of positive-item detection and a few incorrect identifications. Overall the revolver is correctly located and identified in 90.5% of the examples (19/21) with a low false-positive rate of (0.0%, 0/25). Figure 4.21 shows keypoints from the revolver *reference* item superimposed into a baggage item indicating correct identification of the target item in this case.

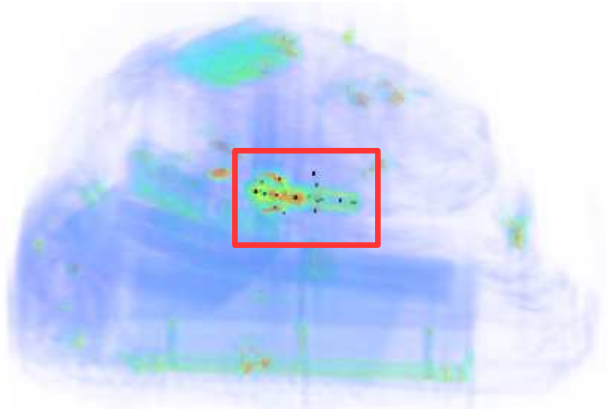
Notably, particular items of interest may be dismantled for concealment in the



(a)



(b)



(c)

Figure 4.21: Correct identification of revolver (x, y, z views)

	Number of bags	Correct Identification	Incorrect Identification
Target Present	21	19 (90.5%)	2 (9.5%)
Clear Bag	25	25 (100.0%)	0 (0.0%)

(a) Revolver confusion matrix

	Number of bags	Correct Identification	Incorrect Identification
Target Present	27	18 (66.7%)	9 (33.3%)
Clear Bag	25	25 (100.0%)	0 (0.0%)

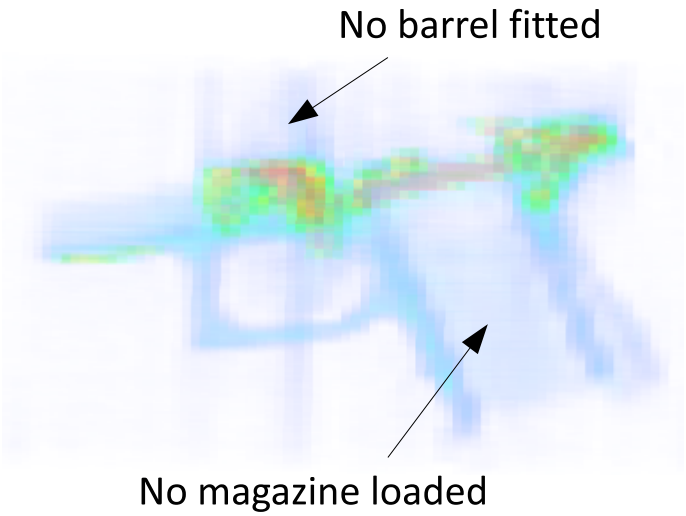
(b) Pistol-frame confusion matrix

Table 4.1: Object-recognition results

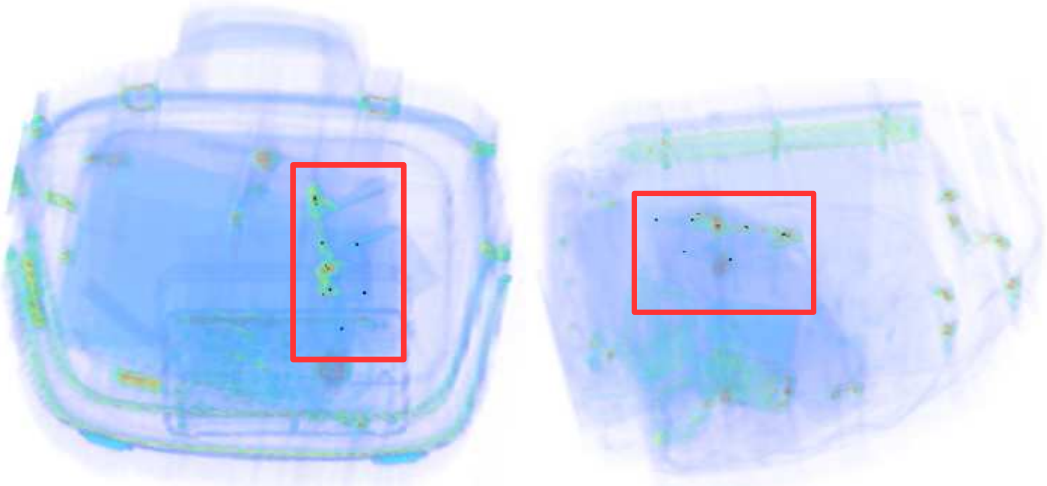
airport-baggage-screening scenario (Shanks and Bradley, 2004). Here we consider a dismantled Glock 9mm pistol with solely its frame (handle and trigger) introduced as the target item (Figure 4.22b). For this example a number of scans were taken (28 with target; 25 negative). This object is mostly plastic with some metal located where the barrel attaches. Consequently we use a lower reference-object-density threshold of $\tau_d = 0.10$ for this target item when forming the verification object (see Figure 4.23). Figure 4.20b shows a histogram of the match metric results for this target from which we can see that a decision threshold is less obvious (than in Figure 4.20a). Taking a threshold value τ_i ($\tau_i = 0.6$) yields the results presented in Table 4.1b where we see this more difficult target correctly located 67% of the time with a low false-positive rate (0.0%). Two examples of correct identification are shown in Figure 4.22c where we can see the pistol frame located in complex baggage. Figure 4.24 shows examples where the frame is incorrectly located, although in both cases the estimated location is close to the actual position of the pistol frame. In some cases the top of the pistol frame is located correctly but in the wrong direction such that the estimated orientation of the pistol is backwards (Figure 4.24a). On other occasions the metallic parts of the pistol frame provide reasonable descriptors but subsequent misalignment occurs and an error in the estimated tilt of the frame occurs (Figure 4.24b). The numerous artefacts within the CT data can be seen to corrupt local gradients around points of interest and this will result in difficulties generating reliable SIFT descriptors. In particular the generation of an invariant coordinate system that copes with variations in orientation will be corrupted by the streak and shadow artefacts.

The lesser performance in this secondary example (pistol frame, Figure 4.22b) can be attributed to the fact that this item is largely made from plastic with a small amount of metal where the pistol slide (barrel) would be attached. Metal artefacts

(a) Disassembled Glock pistol



(b) CT image of pistol frame



(c) Correct location identified in baggage

Figure 4.22: 9mm pistol frame as target

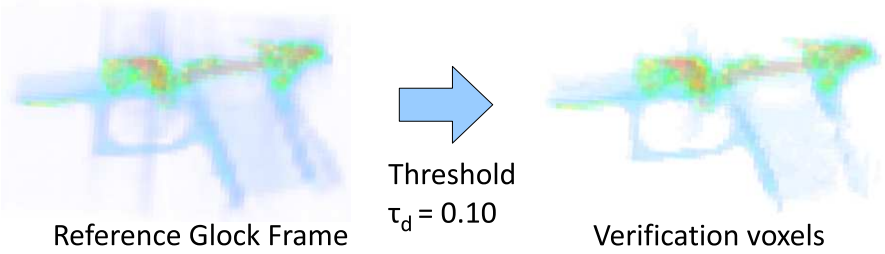
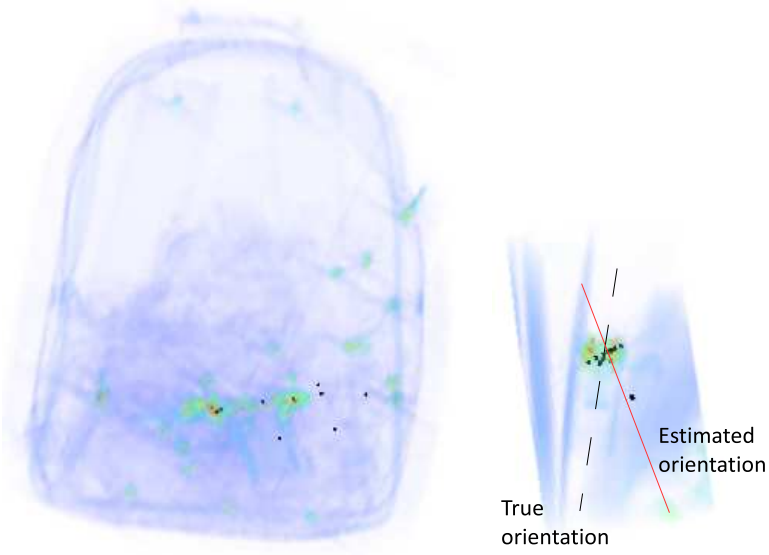
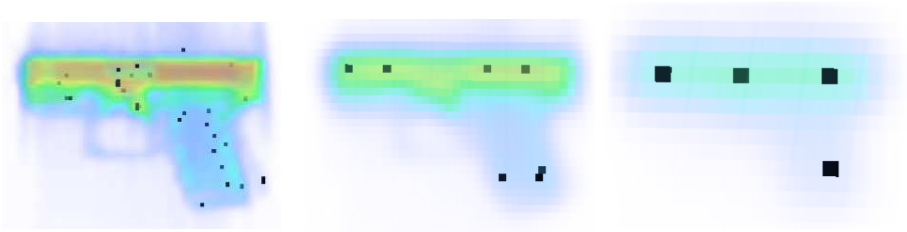


Figure 4.23: Glock frame verification points

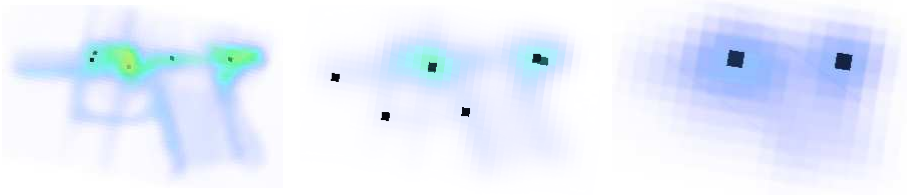


(a) 'Back to front', view from side (b) Incorrect tilt, view from behind

Figure 4.24: Incorrect location of pistol frame



(a) Fully-assembled pistol



(b) Pistol frame

Figure 4.25: Keypoint variation for Glock pistol

that are generated as part of the CT scanning process (Section 3.1.3) can have a similar density to genuine parts of the pistol frame and consequently the 3D image gradients (a key part of the SIFT approach) around points of interest are more easily corrupted by noise. This, coupled with the lower density plastic of the frame, makes matching in this case more complex.

It had been envisaged that the keypoints derived from the frame of the pistol (target item Figure 4.22b) would enable location of a fully-assembled pistol. Experimentally this has proved invalid as a complete pistol has significantly different keypoints in both location and description (Section 4.2) due to the material changes that occur on reassembly - the pistol frame lacks the internal features that would be unaffected when the rest of the pistol is attached at these resolutions. This can be seen in Figure 4.25 where we can see the keypoints derived for the fully-assembled pistol and those for just the pistol frame. Note that the fully assembled pistol contains an empty magazine inside the grip which contains a spring mechanism for loading ammunition. The presence of the magazine results in numerous keypoints in the grip (Figure 4.25a) that are not present in the pistol frame (Figure 4.25b). Also note that the presence of the pistol barrel and slide (large metallic items) in the assembled item removes the pistol-frame keypoints associated with its metallic slide attachment mechanism. This appears to be a generalized problem with the nature of interacting sub-components of any given volumetrically-sampled object interacting and varying the extreme edges, ridges and corners which are key to the SIFT approach.

Additionally, the combined set of data (21 bags containing revolver; 27 bags

Contents	Identification Result		
	Clear	Revolver	Pistol Frame
Clear	25	0	0
Revolver	2	19	0
Pistol Frame	9	0	18

Table 4.2: Confusion matrix of {clear bag, revolver, pistol frame}

containing pistol frame; 25 bags clear) were combined into a single dataset that was processed to identify any cross-related errors of individual item identification. The results of this are represented as a confusion matrix in Table 4.2 where we can see a clear diagonal correlation between the identification of clear bags and of the two targets (revolver/pistol frame) but we can additionally see a difficulty in the identification of the pistol frame. Within aviation screening, the identification of disassembled weaponry (such as a pistol) is considered to be a challenging task for human operators and automatic-recognition algorithms alike (Shanks and Bradley, 2004).

4.5 Conclusions

Our results have shown that the use of 3D SIFT to recognize known objects in complex CT volumes that contain significant metal artefacts and relatively poor resolution is possible with a relative degree of success. The detection of a revolver in complex baggage items shows a high true-positive rate (90.5%) and a low false-positive rate (0.0%) which is a requirement for an airport baggage-screening scenario. However, the relatively poor resolution coupled with its anisotropic nature leads to issues in the identification of smaller items and generalized item sub-parts (Glock 9mm pistol frame, Figure 4.22, Table 4.1b).

In general, this problem could be overcome by considering explicit verification models for the sub-parts of common threat items (e.g. disassembled firearms) or alternatively the consideration of part-based approaches as an extension to the statistical feature-driven recognition of Lowe (2004), as illustrated in Felzenszwalb et al. (2010). Although these part-model approaches (Felzenszwalb et al., 2010) have shown to be successful over general 2D object recognition, it has to be noted that the interaction of volumetrically-sampled sub-components, as shown in this work and its effect subsequently on feature identification, is not one encountered within the 2D object-recognition scenarios considered by Felzenszwalb et al. (2010).

In general, the presence of CT artefacts is thought to be the primary cause

behind false matches in the results presented - the image gradients are corrupted, thus rendering the SIFT gradient histograms subject to a large degree of noise. It is believed that the requirement of the SIFT descriptor to establish an orientation-invariant coordinate set (Section 4.2.4) is prone to error when presented with the CT baggage data, and this can lead to poor matching. Alternative methods of description may suffer less from the imaging artefacts and yield improved matching performance.

The use of a match threshold $\tau_m = 1.2$ in the creation of candidate match set was derived by experimentation. We wish to explore alternative methods in the creation of the candidate match set, in particular the 'distinction' method used by Lowe (2004) of looking for matches that are clearly distinct and less prone to ambiguity.

Overall from our results on the extension of 3D SIFT we identify a number of areas for immediate further investigation:

- Alternative descriptors that are may be more suitable to CT imagery. These descriptors must be orientation invariant.
- Methods of establishing the candidate match set - a fixed threshold has been implemented but we could use the 'distinct' match method of Lowe (2004) as an alternative with the aim of increasing the quality of the matches within the candidate set and at the same time reducing the size of the candidate set to speed the matching process.
- Increasing the number of items used as objects of interest to gain further insight into the recognition task we are investigating.

Chapter 5

Comparison of 3D-feature descriptors

In this chapter we compare the performance of the SIFT descriptor against other descriptors using an extended dataset containing an increased number of target items. We begin with the interest-point-location methodology from previous work (Section 4.2.1). The formulation of a number of 3D interest-point descriptors and a recognition methodology are then described before results and conclusions are presented.

5.1 Introduction

Despite the promising performance of the 3D SIFT descriptor in the detection of known objects (Chapter 4) we are suspicious that its performance is hindered by the quality of the CT data and in particular that the dominant orientation methodology is corrupted by scanning artefacts. We introduce a set of new descriptors ranging from simple to complex, and examine their performance against SIFT. We also examine the selection process in initial descriptor matching to move away from the fixed-Euclidean-distance threshold used in Chapter 4. The number of baggage items being examined is increased, as is the number of target items.

An overview of descriptor generation is shown in Figure 5.1 where we see the separation of interest-point detection from descriptor generation which, in our comparison for this work, can be performed in a number of different ways (as described in Section 5.2). Interest-point locations for an input volume are generated using the SIFT-derived methodology described in Section 4.2.1. Descriptors for each volume are generated using these locations. The location of the keypoint is stored as part of the descriptor to facilitate a relative-position-consistency check in a subsequent recognition methodology (Section 5.3).

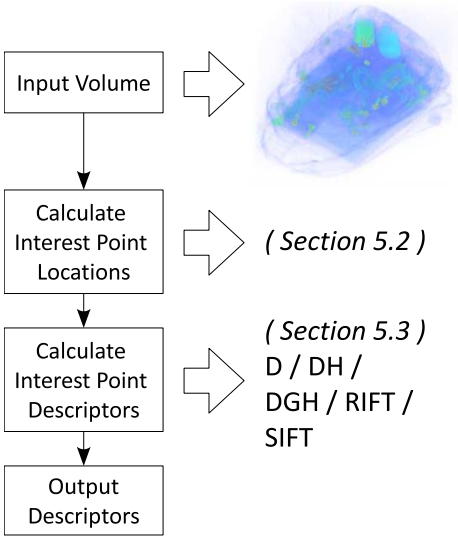


Figure 5.1: Descriptor generation

5.2 3D point of interest descriptors

We now wish to characterize the local neighbourhood. We give details of a range of approaches for this characterization with increasing levels of complexity, from a simple local density average through density and gradient histograms, leading on to 3D extensions of RIFT (Lazebnik et al., 2005) and SIFT (Lowe, 2004). All descriptors need to be invariant to orientation: for SIFT, this is achieved through determination and re-orientation around a 'dominant' orientation. In other cases, orientation invariance can be achieved by ensuring that the description methodology does not vary as the azimuth/elevation subtended by contributing voxels is considered.

5.2.1 Local-point-of-interest-neighbourhood function

In determining the location of points of interest within a volumetric image we use the Difference-of-Gaussians methodology described in Chapter 4, Section 4.2.1. The use of a consistent methodology for keypoint location allows us to compare the effect of different descriptors on the detection of items of interest.

Following from the identification of interest-point locale, we now define a localized neighbourhood function, extending this from earlier work in 2D (Lowe, 2004).

A Gaussian-window function, $w(d, \sigma)$, is used to limit the contribution of voxels around the point of interest to those in the local neighbourhood:

$$w(d, \sigma) = exp\left[-\left(\frac{d}{\sigma}\right)^2\right] \tag{5.1}$$

where d is the voxel distance from the point of interest to the contributing voxel and σ is used to determine the extent of the local contribution. The way that this function is used is given with each of the descriptor formulations. It should be noted that, given the definition of distance in voxel units, this window will remain consistent with the resolution of the volume being examined. Values chosen for σ are stated in the following section as each descriptor is discussed.

5.2.2 Simple density descriptor (D)

This density descriptor is a simple Gaussian average around the point of interest, P, defined in Equation (5.2):

$$D_P = \frac{\sum_k I_k w(d_k, \sigma)}{\sum_k w(d_k, \sigma)}, \tag{5.2}$$

for voxel k, with a density I_k and a voxel distance d_k from the interest-point location. The local neighbourhood function, $w(d_k, \sigma)$, is as defined in Section 5.2.1. This descriptor is orientation invariant as we consider only the Euclidean distance from the keypoint to the contributing voxel.

This is a simple descriptor and is included as a baseline comparison to its more complex counterparts.

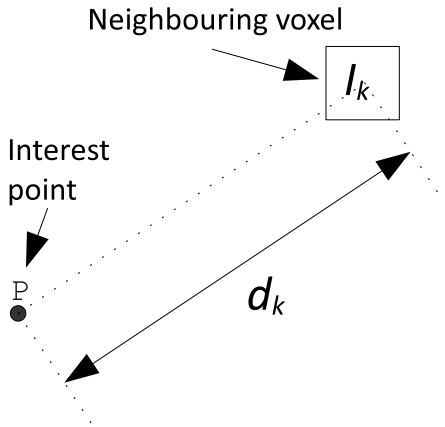
5.2.3 Density-histogram descriptor (DH)

By contrast, this second descriptor defines the local density variation at a given interest point as a weighted histogram over a continuous density range. The density range is $[-1.0, 2.0]$, in line with the resampled cubic-voxel volume, and is split into N_{DH} bins resulting in a bin width (β_{DH}) of $3.0/N_{DH}$. The voxel density for point k is I_k and this is used to determine which histogram bin is active. Given the local area function $w(d_k, \sigma)$, defined in Section 5.2.1, an addition of $w(d_k, \sigma)$ is made to the appropriate histogram bin where d_k is the voxel distance from the point of interest to voxel k. The weighted-density histogram for point of interest P is:

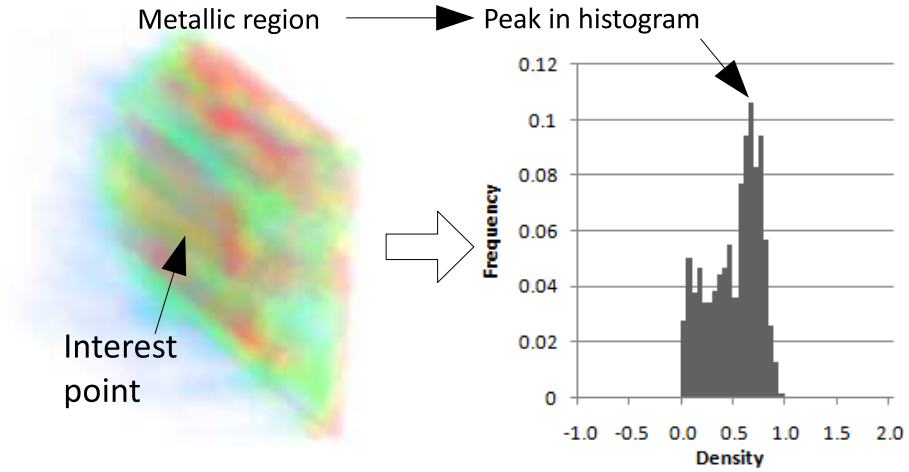
$$DH_P(i) = \sum_k \begin{cases} w(d_k, \sigma) & \text{if } (|I_k - I_i| < \beta_{DH}/2) \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

where I_i is the central density for the i^{th} histogram bin.

The descriptor is normalized to unity area on completion. Figure 5.2a shows



(a) Local neighbourhood voxels



(b) Example density-histogram descriptor

Figure 5.2: Density-histogram calculation

an example point of interest, P, with one of its neighbouring voxels of density I_k. Figure 5.2b shows an example of a density histogram derived from an interest point that is located near a metallic region. It can be seen from this that the resulting density histogram has a peak due to the high concentration of metal within the neighbourhood. Again, orientation invariance is maintained as no reference is made to relative voxel positions in the formulation.

5.2.4 Density-gradient histogram descriptor (DGH)

In a variant of the previous descriptor, here we calculate the density-*gradient* magnitude in the neighbourhood of the interest point and then accumulate these in a histogram. The density-gradient magnitude is calculated for all voxels in the volume using a central-difference formulation to ensure that the gradient location is aligned with the voxel grid. The density-gradient-magnitude range is [0.0, 4.0] (given a

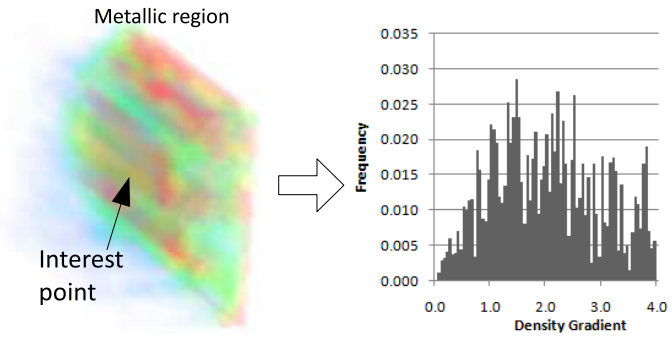


Figure 5.3: Density-gradient-histogram calculation

voxel dimension of $2.5mm$, the vast majority of gradient values lie below 4.0) and is divided into N_{DGH} bins, resulting in each bin having a width (β_{DGH}) of $4.0/N_{DGH}$. The voxel-gradient magnitude for voxel k is δ_k and this is used to determine which histogram bin is active. Once the active histogram bin is determined, an addition of $w(d_k, \sigma)$ is made to the corresponding histogram entry, with $w(d_k, \sigma)$ again defined in Section 5.2.1.

$$DGH_P(i) = \sum_{i=1}^{N} \begin{cases} w(d_k, \sigma) & \text{if } (|\delta_k - \delta_i| < \beta_{DGH}/2) \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

where δ_i is the central density gradient for the i^{th} histogram bin.

The descriptor is normalized to unity area on completion, as before. It is notable here that, as objects under consideration for detection can occur in any orientation, the gradient *magnitude* is used rather than the gradient-*orientation* approach, frequently used for recognition tasks in 2D (Dalal and Triggs, 2005).

Figure 5.3 shows the same point of interest as for Figure 5.2b but now with the density-gradient histogram. It is not as obvious how the histogram relates to the volume given the noise.

5.2.5 Rotation invariant feature transform (RIFT)

Lazebnik et al. (2005) developed the Rotation Invariant Feature Transform (RIFT). The RIFT descriptor examines the local neighbourhood gradients with reference to a radial vector emanating from the point of interest. Histograms are constructed from the gradient orientation and magnitude. Multiple histograms are derived following segmentation of the local neighbourhood into a series of rings centred on the point of interest. RIFT has been shown to operate well in standard 2D imagery, and is used in our work as it is more complex than the simple histograms described above, but is not as complex as the SIFT descriptor (Lazebnik et al., 2005; Lowe, 2004).

Before describing our extension of RIFT to 3D we consider our variant in 2D.

Figure 5.4a shows a point of interest, P, and neighbouring region. For each neighbouring pixel, I, a unit vector in the direction **PI** is calculated: $\mathbf{R_I}$. The gradient at pixel I is $\mathbf{g}_I$. The angle between the gradient and radial vector is θ_I. A weighted histogram is constructed based on values of θ_I in the range $[-\pi, \pi]$. There are N_b bins in this histogram representing angular regions $2\pi/N_b$ radians in size. For each gradient and angle, an addition to the histogram of $|\mathbf{g}_I| .w(d_I, \sigma)$ is made as shown in Figure 5.4a. Note again that the function $w(d_I, \sigma)$ limits the contribution to the local neighbourhood. In addition to the histogram, N_r rings, of width d_w pixels, are also defined as shown in Figure 5.4b (with $N_r = 3$ as an example). One histogram is generated for each region and each histogram is normalized by the area of its ring to prevent bias to regions of greater area. The complete descriptor histogram is normalized to unity area. The resultant descriptor has $N_r \times N_b$ elements.

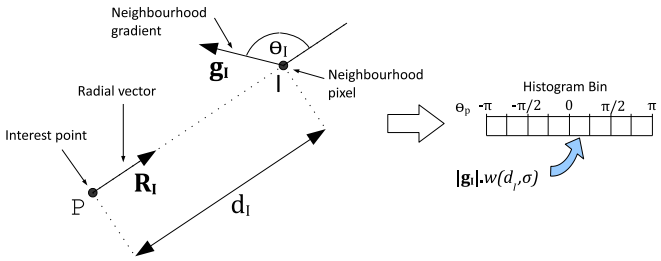
The extension of this descriptor to 3D is straightforward noting that, due to rotation symmetry in 3D, the radial histograms only cover values of θ_p in the range $[0, \pi]$ and the normalizations refer to region *volumes* rather than areas. One additional normalization is required in the move to 3D: the histogram summations are normalized by bin surface area to remove bias towards equatorial bins. Figure 5.5 shows an example with 4 bins per histogram: bins A, B, C and D. If the volume has unit radius, bins A and D have a surface area of $\pi\left(2 - \sqrt{2}\right)$, whereas bins B and C have an area of $\pi\sqrt{2}$. These areas are used to normalize the summations for each bin. This step is not required in the 2D case as all histogram bins have the same sector angle.

As with other descriptors, the final step is to normalize the complete descriptor to sum to unity.

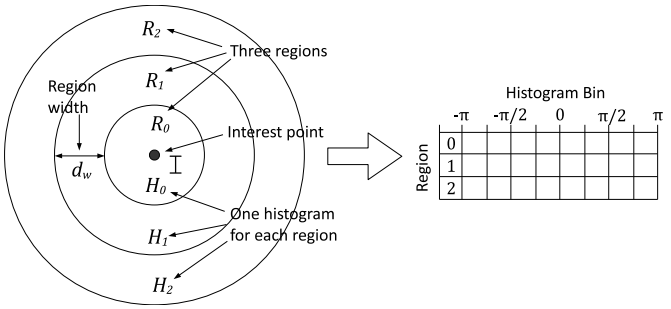
Figure 5.6 shows the RIFT descriptor generated for the same metallic region as used in the density histogram and density-gradient histogram explanations (Figures 5.2b/5.3). This plot shows that, for this example, the gradients tend to point toward to point of interest rather than away, which is expected as it is a high density (metallic) region.

5.2.6 3D scale invariant feature transform (SIFT)

We use the SIFT descriptor as derived in Chapter 4 in our comparison. The SIFT parameters are the same as before: $N_g = 3$ and $N_v = 3$ (Section 4.2.5); $R_{max} = 9$ for the Gaussian weighting (Section 4.2.4) resulting in a descriptor that characterizes a $9 \times 9 \times 9$ voxel volume.



(a) 2D radial geometry



(b) 2D radial regions

Figure 5.4: 2D-RIFT descriptor

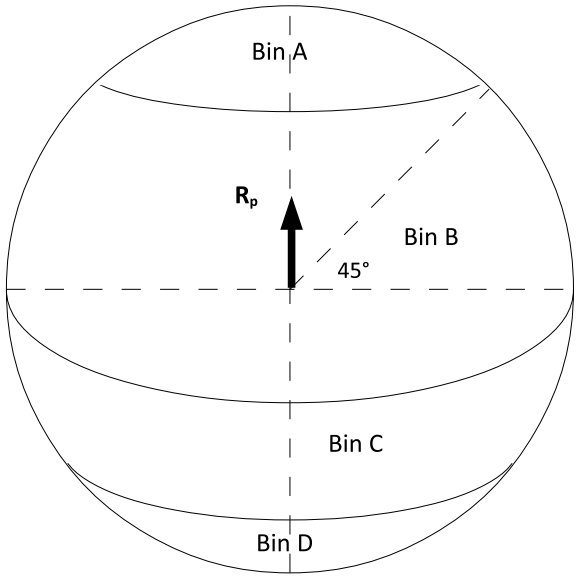


Figure 5.5: 3D RIFT-bin normalization

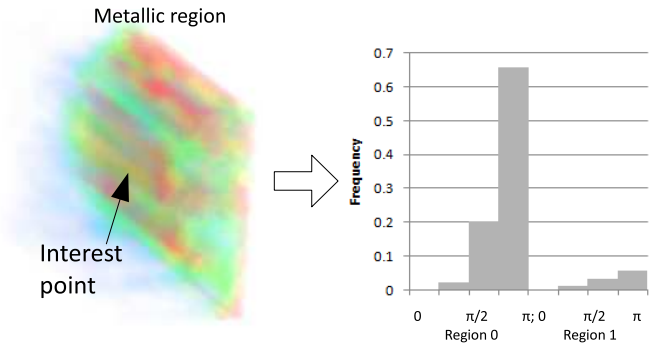


Figure 5.6: RIFT descriptor example

5.3 Object-detection methodology

An object-detection-system methodology is shown in Figure 5.7. Here we start with a known reference item from which descriptors are calculated. A candidate baggage item is received and processed to determine its descriptors. The matches between the reference and candidate item are filtered in an attempt to remove false matches. The output set of matches from this process are referred to as the correspondence set.

Two methods are used when forming the correspondence set:

a) The method of Lowe (2004) is employed where a match is accepted to the correspondence set if the ratio of the first to second best match distances is less than 0.8. We refer to this method as the distinction method. We consider this process from the candidate to the reference, i.e. a candidate/reference pair is added to the correspondence set if it is distinct compared to matches between the same candidate and the other reference descriptors.

b) We sort the matches by Euclidean distance in ascending order. We then choose a fixed percentage of the best matches as the correspondence set. We refer to this method as the percentile method, with parameter p defining the percentage of matches used. The chosen value of p is determined through experimentation: too small a value will result in few matches and limited recognition performance; too large a value will result in many poor quality matches in the correspondence set, which will increase the time taken to recognize an object.

Given the large number of possible false matches in this formulation we make use of RANSAC (Fischler and Bolles, 1981) to find an optimal match using the correspondence set as the input. This RANSAC formulation is used to select a set of three possible matches from the correspondence set from which a 3D transformation is derived using a SVD approach (Arun et al., 1987) as in Chapter 4. Following estimation of the transformation we check to see if the three RANSAC-selected

matches are consistent: the reference set and candidate set should have similar shapes: relative distance errors should be less than ϵ_r ($\epsilon_r = 10mm$). It should be noted that the one-to-one relationship between voxel measurements and real-world distances allows the tolerance ϵ_r to be specified in real-world measurements (i.e., mm). This constraint aids the matching process by quickly rejecting poor quality selections prior to the verification stage.

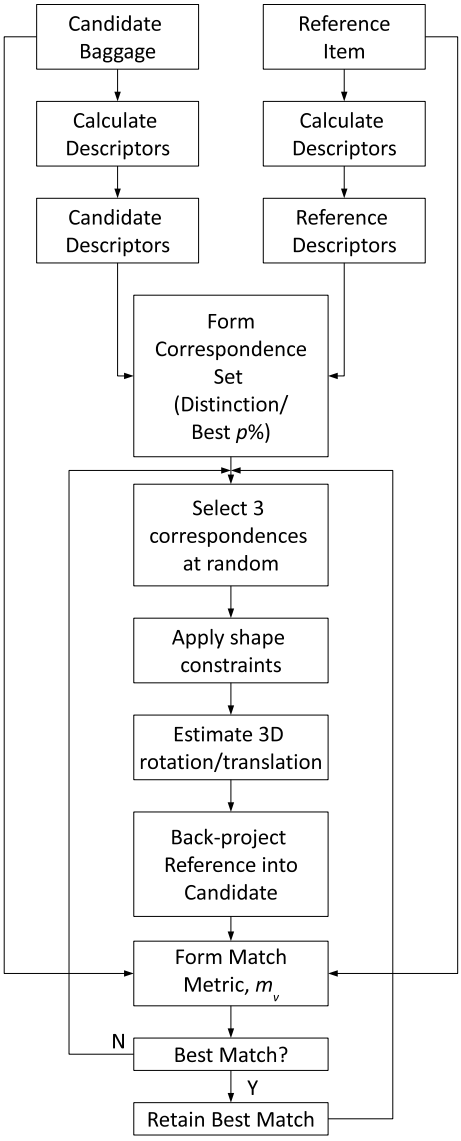


Figure 5.7: Object-recognition methodology

If the relative distance criterion is passed, a secondary verification is performed. Locations in the reference object with a density above a threshold τ_d ($\tau_d = 0.15$) are recorded to form a set of density-verification locations. The threshold is applied in order to reduce the number of low-density artefacts in the verification stage. The verification locations are transformed into the candidate baggage-item space

using the least-squares estimate of rotation and translation provided by the SVD formulation. Given N_v verification points we then form a quality of match metric, m_v, by examining the density differences between the verification locations in the reference item and the candidate baggage item:

$$m_v = \frac{\sum_{k=1}^{N_v} |I_k - \psi_k|}{\sum_{k=1}^{N_v} \psi_k}, \tag{5.5}$$

where ψ_k is the density at the k^{th} verification point in the *reference* item and I_k is the density of the voxel closest to the k^{th} transformed verification point in the *candidate* baggage item. The measure is normalized by the sum of the densities of the verification points in the reference item, as shown, to provide a measure that does not vary too greatly between different reference items.

The set of descriptors for comparison, described in Section 5.2, were computed using the parameter settings shown in Table 5.1 which were derived from experimentation. As an example, when too few histogram bins are used for the density histogram or density-gradient histogram the resultant descriptors are not distinctive, yielding too many matches, whereas if there are too many bins used then the descriptors are too unique and there are not enough quality matches. Both situations yield poor matching results. The results of this comparison using the proposed object-detection methodology and the parameter settings listed in Table 5.1 are presented in the next section.

Descriptor	Settings	Elements per Descriptor
Density	$\sigma = 1.0$	1
Density Histogram	$\sigma = 3.0$, $N_{dh} = 60$	60
Density Gradient Magnitude Histogram	$\sigma = 3.0$, $N_{gh} = 80$	80
RIFT	$\sigma = 3.0$, $N_b = 4$, $N_r = 2$, $d_w = 3.0$	8
SIFT	$N_g = 3$, $N_v = 3$, $N_a = 8$, $N_e = 4$	864

Table 5.1: Descriptor settings

5.4 Results

5.4.1 Experimental program

We begin by examining the distinction methodology when forming the correspondence set for each descriptor and target item (Section 5.4.2). An examination of reference-item orientation when originally scanned is then made with the aim of improving recognition performance (Section 5.4.4). We then investigate a recognition system that uses a fixed percentile correspondence-set approach and demonstrates superior performance over that obtained with the distinction methodology (Section 5.4.5).

Statistical results are presented using true-positive and false-positive rates. An explanation of this approach is given in Appendix C.

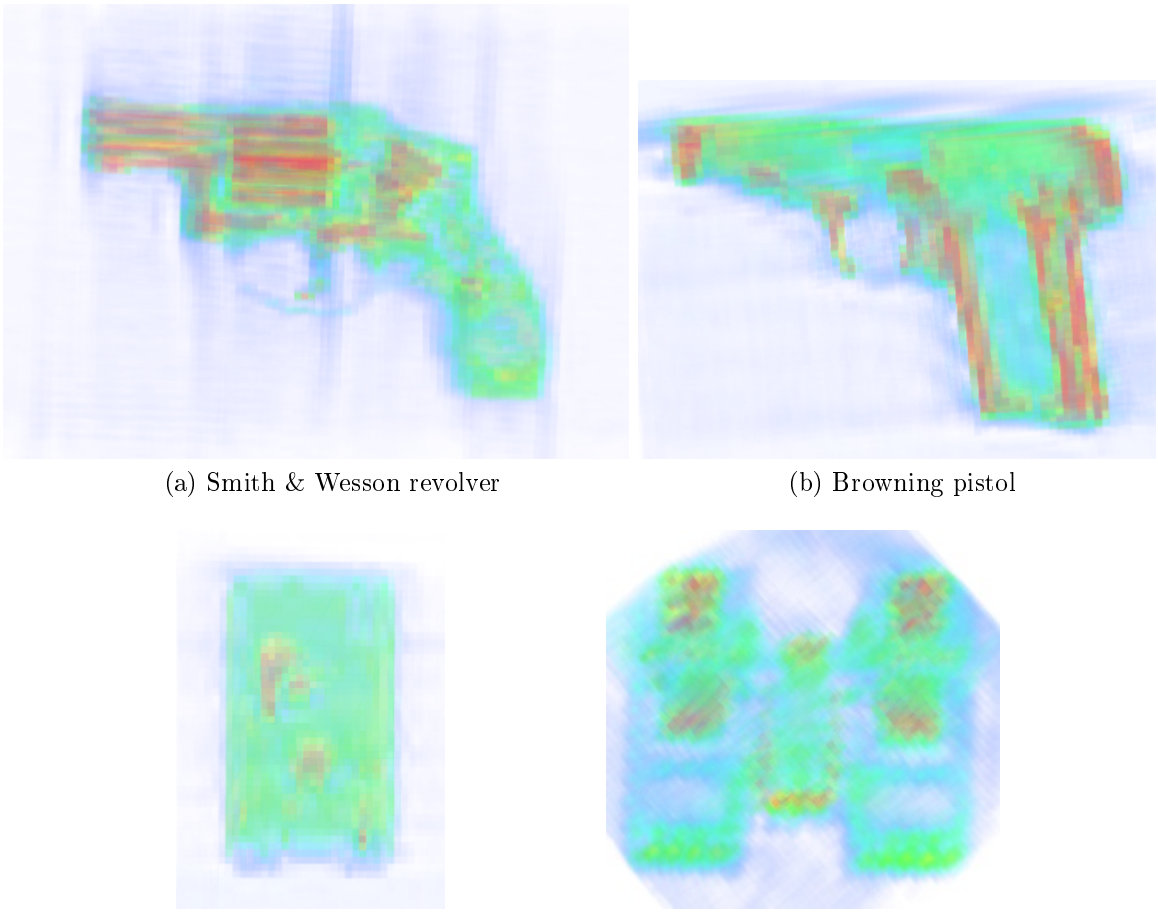
5.4.2 Distinction-methodology-correspondence set

First we consider the distinction method when forming the correspondence set which is true to Lowe (2004) rather than the fixed-threshold approach that was successfully used in Flitton et al. (2010).

For this comparative study four target items of interest were used (Smith & Wesson revolver; Browning pistol; Apple iPod; compact binoculars) scans of which are shown in Figure 5.8. Furthermore a mix of baggage types were scanned (e.g. holdalls, suitcases, handbags) containing a variety of clutter items as would be found in a typical airport scenario, including and excluding these items of interest. Table 5.2 shows the number of baggage items scanned which contained one of these target items or which were left clear of the named targets but still contained regular background clutter. Note that the number of clutter-only bags has increased significantly from the initial SIFT experiments in Chapter 4.

Baggage item contents	Scans in collection
Smith & Wesson revolver + clutter	21
Browning pistol + clutter	30
Apple iPod + clutter	15
Compact binoculars + clutter	14
Clutter only	180

Table 5.2: Items scanned



(a) Smith & Wesson revolver

(b) Browning pistol

(c) Apple iPod

(d) Binoculars

Figure 5.8: Reference CT object volumes used for detection

CT scans of each baggage item are analyzed using the object-detection methodology outlined in Section 5.3. From this, each baggage item produces a verification match metric result, m_v, as described in Section 5.3 (a measure of similarity between the reference item and the baggage item). A decision on whether a target item has been detected is made by comparing the verification match metric result, m_v, against a detection threshold, τ_m. Given that we know which baggage items contain the target items and which do not, we can calculate both a true-positive de-

tection rate, $TP(\tau_m)$, and a false-positive detection rate, $FP(\tau_m)$, for a given setting of τ_m. Our analysis uses Receiver-Operating-Characteristic (ROC) plots (Fawcett, 2006) to investigate the overall system performance as each descriptor type is used. These plots show $TP(\tau_m)$ against $FP(\tau_m)$ and indicate the trade off between true detection of threat items versus false detection as the detection threshold, τ_m, is varied. When producing a numerical performance result we choose to quote the true-positive rate for minimal false-positive rate ($<1\%$) rather than the ROC equal error rate (Schuckers, 2010) as we feel that this is more applicable to the operating conditions of such a system in an operational security-environment (even a moderate false-positive rate is not desirable).

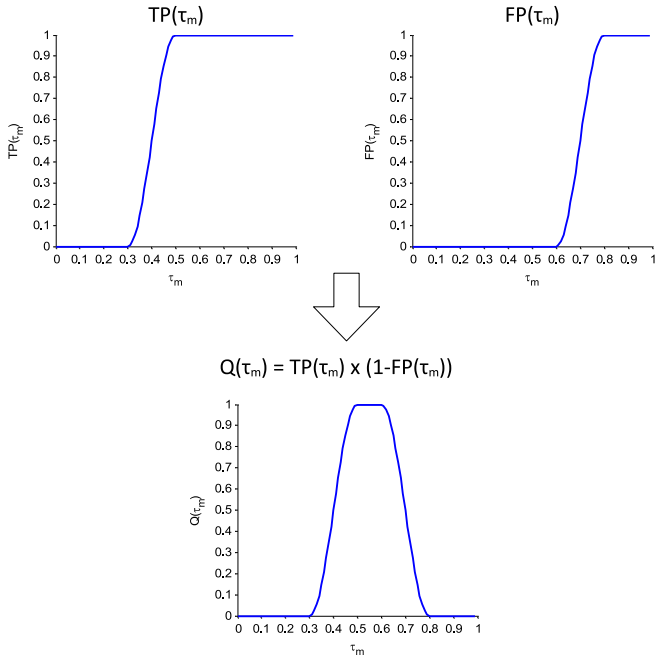
The ROC plot gives one aspect of performance. We also form a plot that shows a measure of tolerance to error given the value of the detection threshold, τ_m, should a fixed value be chosen to decide the presence of the target item. We refer to this as the *threshold quality*, $Q(\tau_m)$, where:

$$Q(\tau_m) = TP(\tau_m)\,[1 - FP(\tau_m)] \tag{5.6}$$

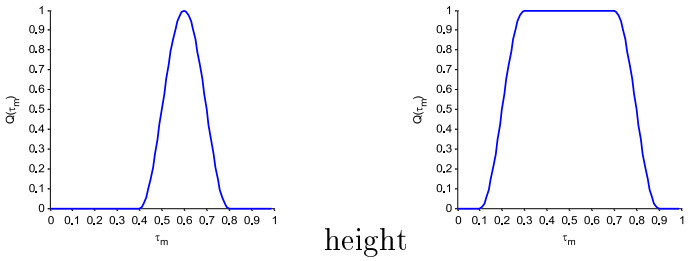
Figure 5.9a shows how the true-positive and false-positive rates are combined to form the threshold quality. The width of the threshold quality plot indicates the separation between the rise in true-positive rate and the rise in false-positive rate. The maximum value of the threshold quality peak is also indicative of performance. If the true-positive and false-positive rates are well separated then the threshold quality will reach a peak value of 1.0 which would indicate a perfect ROC plot. However, if the true-positive and false-positive-transition regions overlap, the threshold-quality peak will be less than 1.0. Figures 5.9b and 5.9c show threshold-quality plots for two systems, both with perfect ROC plots. It can be seen in Figure 5.9b that the threshold-quality peak is narrow indicating that the true-positive transition region is close to the false-positive transition region. A better scenario is shown in Figure 5.9c where the threshold-quality peak is broad indicating a large separation between the true-positive and false-positive transition regions. This broad peak indicates that, when allocating a value to the detection threshold (τ_m), a greater tolerance to error in its assignment exists.

5.4.3 Distinction methodology results

Results of this work are presented as ROC plots using the legend given in Table 5.3. We begin with an investigation of detection performance using the distinction approach of Lowe (2004) to form the correspondence set and then we look at the



(a) Threshold quality derivation



(b) Poor threshold quality

height

(c) Good threshold quality

Figure 5.9: Threshold quality

Descriptor	Legend
Scale invariant feature transform	SIFT
Density	D
Density histogram	DH
Density-gradient histogram	DGH
Rotation invariant feature transform	RIFT

Table 5.3: Plot legend

performance using the percentile method proposed in our earlier work (Flitton et al., 2010).

ROC plots for detection of the revolver, pistol, iPod and binoculars when using the distinction method are shown in Figure 5.10. It can be seen that there is a considerable variation in detection performance between the descriptor types, as well as differing levels of detection of each target item.

For the revolver (Figure 5.10a) the best result using the distinction method is obtained using the RIFT descriptor with a detection rate of ∼95% with detection rates using D, DH and DGH at ∼60/70% . The performance of SIFT is poor with a detection rate of ∼20%.

The pistol performance is poorer (Figure 5.10b) with a detection rate of ∼55% with a negligible false-positive rate using DGH. This is closely followed by D and DH descriptors (∼50%) with RIFT and SIFT both poor (∼20%).

The iPod performance is worst (Figure 5.10c) with a detection rate of ∼20% using the RIFT descriptor, closely followed by D, DH and DGH (∼15%) with SIFT again the worst performing (∼5%).

Detection of the binoculars is ∼80% (Figure 5.10d) with negligible false positives using DGH. Detection using RIFT, D and DH descriptors is ∼50% with SIFT again worst with a detection rate of ∼20%.

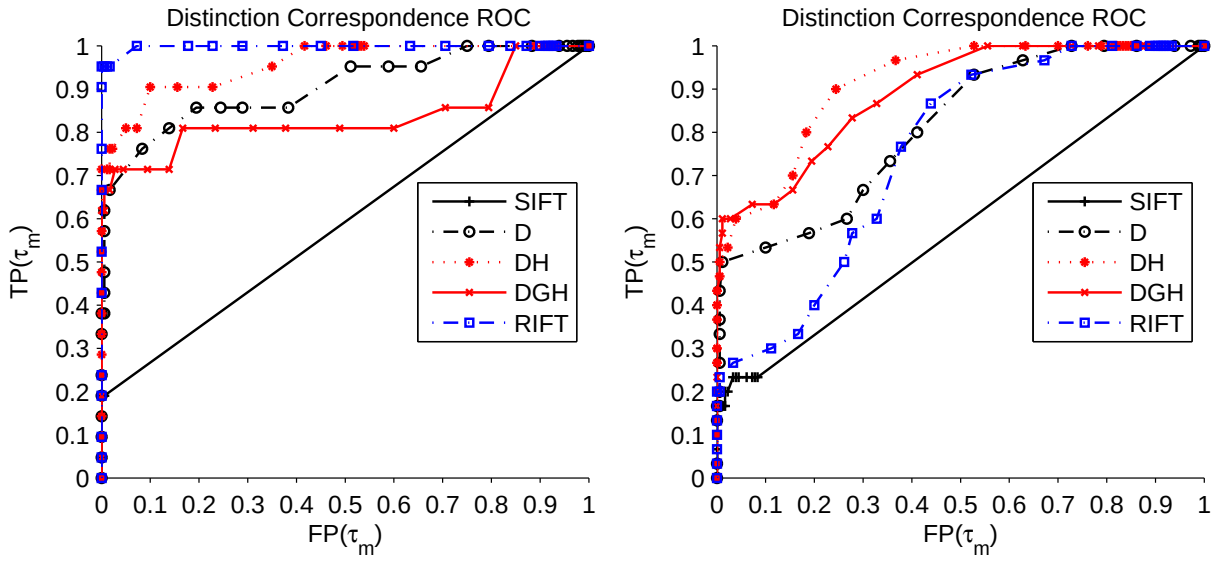
5.4.4 Examination of reference item orientation

Two immediate questions arise from further consideration of these results:

a) why is the pistol-detection rate (∼55%) poorer than the revolver (∼95%) given that they are similar items in both size and density characteristic?

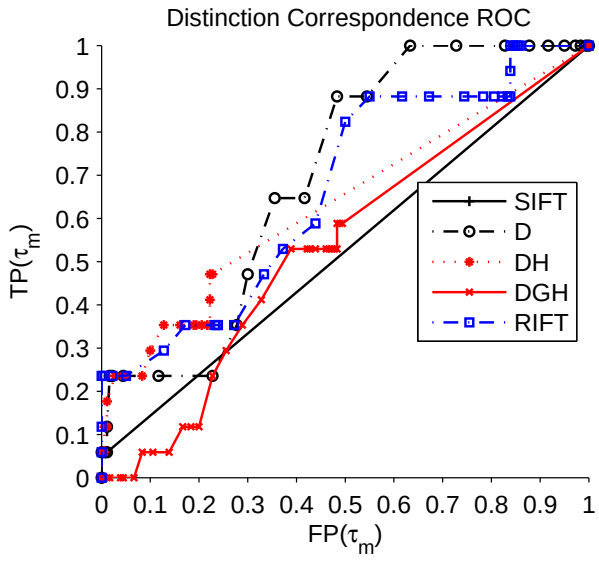
b) why does the use of the SIFT descriptor yield much poorer results when compared to simpler descriptor types?

An investigation into the poor quality of the pistol results compared to those of the revolver indicated that the scan quality of the reference item affects performance.

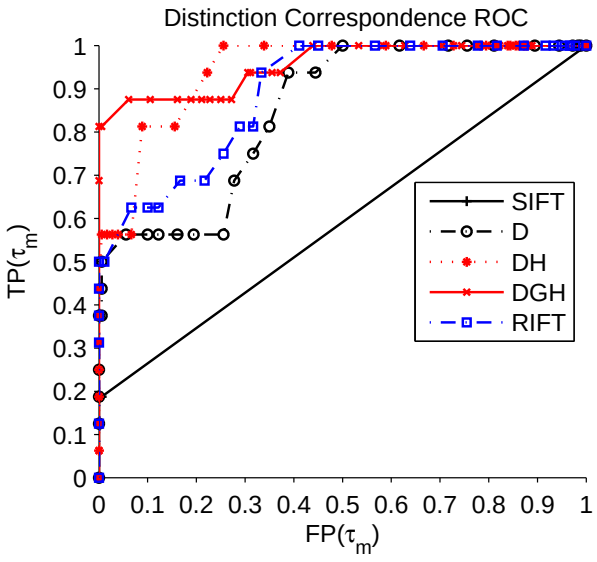


(a) Revolver

(b) Pistol



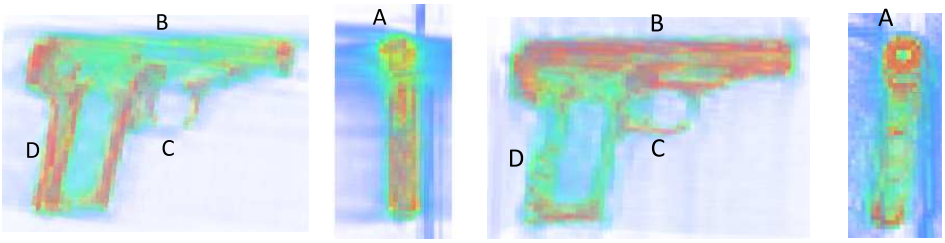
(c) Apple iPod



(d) Binoculars

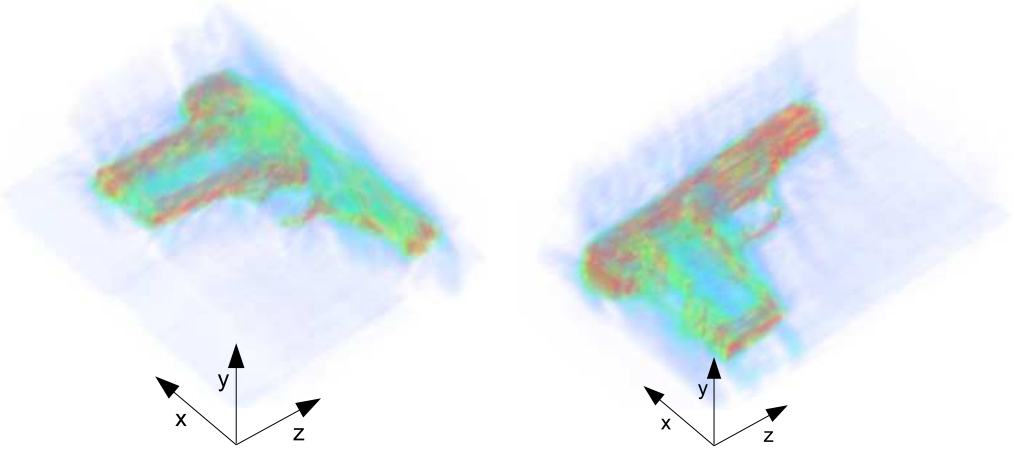
Figure 5.10: Target item ROC curves using distinction to form correspondence set

Figure 5.11a shows the reference used to create the results in Figure 5.10b. Figure 5.11b shows an alternate reference scan of the same Browning pistol. Note in this secondary example (Figure 5.11b) the clarity of the pistol muzzle (A) compared to Figure 5.11a. Also note apparent density differences in the barrel (B), trigger guard (C) and grip (D) caused by metal artefacts and anisotropic scanning. These differences will affect the resulting descriptors, both in value and location, and this has obvious implications for location of similar points in randomly-scanned baggage items. The difference between these scans is the orientation of the pistol relative to the CT scanner z axis, as shown in Figure 5.12. The original pistol reference (Figure 5.12a) was orientated such that the barrel was orthogonal to the z axis resulting in the barrel cross-section being scanned with a 5mm resolution (the CT-slice spacing, see Section 3.1). The alternate pistol reference (Figure 5.12b) was scanned such that the barrel was parallel to the z axis resulting in a barrel cross-section-pixel resolution of $\sim$1.6mm (the slice-pixel resolution - see Section 3.1) and hence greater muzzle clarity.



(a) First reference (b) Second reference

Figure 5.11: Browning pistol reference-item quality



(a) First reference item (b) Second reference item

Figure 5.12: Browning reference-item orientation in CT-baggage scanner

Figure 5.13 shows the ROC plot using match distinction to form the correspondence set when using the alternate pistol reference. Here we can see a better detection rate of ∼85% using DH descriptor (up from ∼50%). The RIFT descriptor has a detection rate of ∼70% (up from ∼20%) with DGH at ∼60% (from ∼55%), density at ∼50% (unchanged) and SIFT at ∼20% (unchanged).

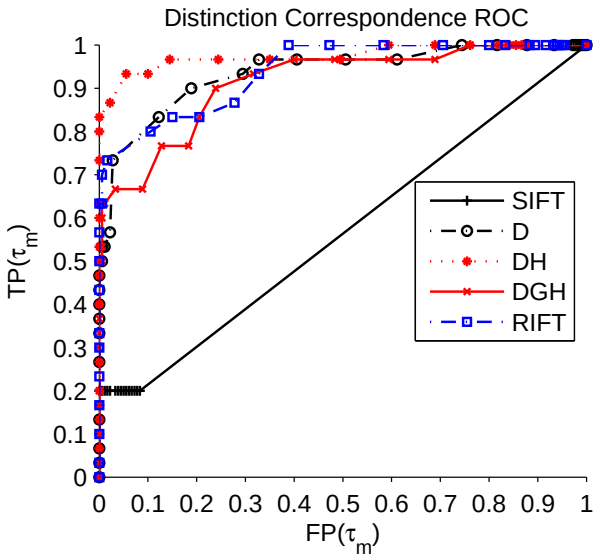


Figure 5.13: ROC using second Browning pistol as reference

We combined the results for both pistol references by choosing the result with the lowest verification match metric value, m_v, to observe if the combination would provide increased levels of performance. Figure 5.14 shows the ROC plot for this situation where we can see that an improvement does occur (compared to the individual reference item results shown in Figure 5.10b and Figure 5.13). The best performance again comes from DH with a detection rate of ∼90% with negligible false positives (up from ∼85%). The performance using the other descriptors is also improved: Density ∼75% (up from ∼50%); DGH at ∼80% (up from ∼60%); RIFT up slightly at ∼75% (from ∼70%); SIFT at ∼35% (up from ∼20%).

5.4.5 Fixed-percentile-correspondence set

An investigation into why the use of the SIFT descriptor yielded poor detection results was carried out. Analysis of the correspondence set showed that, when using match distinction, very few of the SIFT matches were deemed suitable. Table 5.4 shows the mean correspondence-set size (as a % of total matches) for each target item and each descriptor when analyzed over the datasets given in Table 5.2. For D, DH, DGH and RIFT descriptors we see correspondence-set sizes between 0.80% and

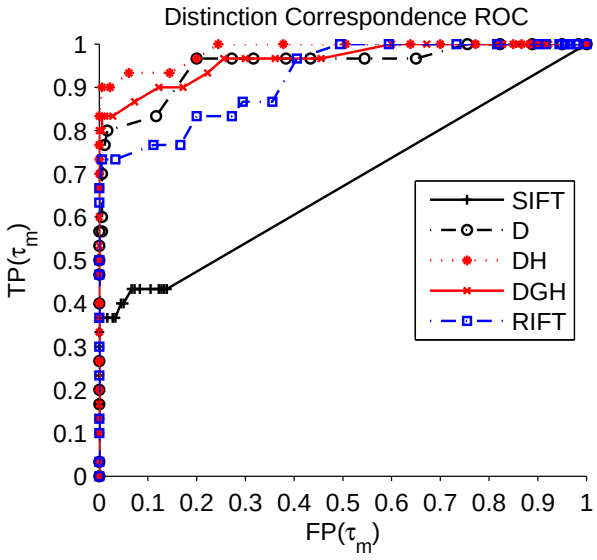


Figure 5.14: ROC for combination of pistol results

Descriptor	Revolver	Pistol	iPod	Binoculars
Density	2.31 ± 0.10	2.47 ± 0.07	3.08 ± 0.10	1.71 ± 0.11
Density Histogram	0.80 ± 0.31	1.32 ± 0.30	1.20 ± 0.50	0.96 ± 0.27
Density-gradient histogram	1.55 ± 0.23	1.18 ± 0.23	0.93 ± 0.20	0.81 ± 0.18
RIFT	1.39 ± 0.14	1.05 ± 0.15	1.17 ± 0.20	1.15 ± 0.11
SIFT	0.02 ± 0.01	0.07 ± 0.06	0.02 ± 0.01	0.01 ± 0.01

Table 5.4: Mean correspondence-set size (as % of total matches) using distinction methodology over set of items in Table 5.2

3.08% of the total number of matches. When compared to these descriptors, the SIFT descriptor has very few matches in the correspondence set: between 0.01% and 0.07%. This is indicative of poor quality descriptors (very few pass the distinction criterion) and it would appear that this restricts its performance: true matches are rejected from the correspondence set and not enough are made available to the object-detection method for reliable recognition of the target items.

It is notable that the use of the distinction method differs from the selection method used in our initial work (Chapter 4) where significantly improved SIFT-3D-object-detection results were obtained when using a fixed selection threshold.

In light of these results and with the support of the earlier work (Flitton et al., 2010) we vary the method used to form the correspondence set away from the sem-

inal 2D-SIFT variation (Lowe, 2004) and use our alternative percentile method as previously discussed in Section 5.3. Rather than using the match-distinction method we instead sort the matches by match distance and then choose a fixed percentage of the best matches.

Figure 5.15 shows the results when the best 2% of matches are chosen to form the correspondence set.

For the revolver (Figure 5.15a) we can see near 100% detection with minimal false positives using DH, DGH and RIFT descriptors. Both Density and SIFT descriptors have detection rates ∼85%.

Using the second pistol reference (Figure 5.15b) we again see near 100% detection using the RIFT descriptor, closely followed by DH and DGH (∼90%) with SIFT at ∼65% and Density at ∼35%.

The iPod detection is still poor (Figure 5.15c), though slightly improved, at ∼30% (increased from ∼20%) using DH, followed by DGH, RIFT and SIFT at ∼20%. The density descriptor has a detection rate of ∼0% using our negligible false-positives-detection-rate definition.

The binoculars show near 100% detection (Figure 5.15d) using RIFT, DGH and SIFT, with DH close behind at ∼95%. The density descriptor is again poor with a detection rate of ∼0%.

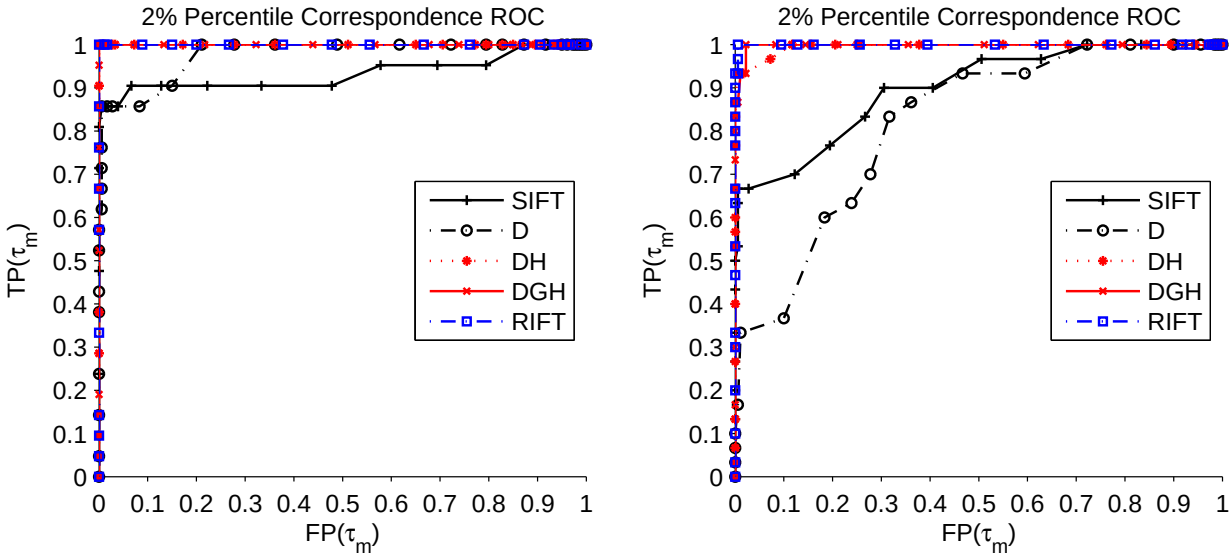
Given a number of ROC plots that appear to show 100% detection rates, mainly due to the limited amount of target items, we can also investigate performance using the threshold quality, $Q(\tau_m)$, as the detection threshold, τ_m, is varied (Equation (5.6)). threshold-quality plots relating the the ROC plots in Figure 5.15 are given in Figure 5.16.

Figure 5.16a shows the plot in the case of the revolver where we see the superior performance of the DH descriptor and RIFT descriptor over the DGH descriptor that it is not possible to see in the ROC plots (Figure 5.15a). Both the DH and RIFT descriptor reach a peak when $\tau_m \simeq 0.45$ and then fall off when $\tau_m \simeq 0.6$. The DGH only reaches a peak for $\tau_m \simeq 0.55$ and then almost immediately starts to fall away. The implication for this, in a noisy environment, would be that the DH and RIFT descriptors would be more reliable than the DGH descriptor.

Figure 5.16b shows the threshold quality for the second pistol reference. Here we see that, although both the RIFT and DGH descriptors reach a peak of 1.0, they quickly fall away. This does not appear to be as good as the revolver.

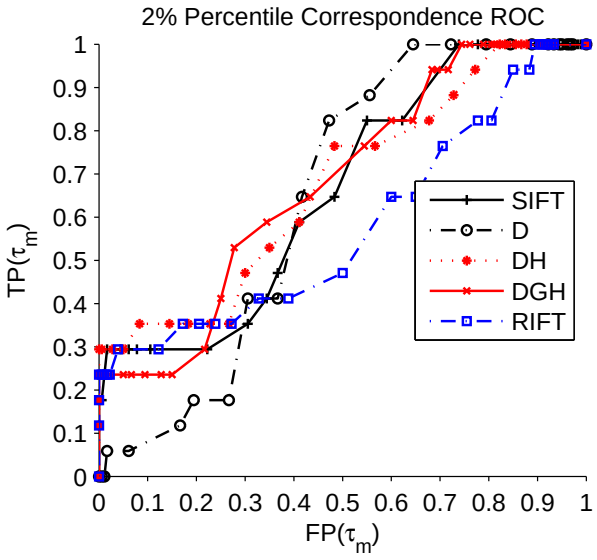
Figure 5.16c shows the results for the Apple iPod. Here we see poor results already indicated by the ROC plot (Figure 5.15c).

Figure 5.16d shows the results for the binoculars. Here we can see that the RIFT

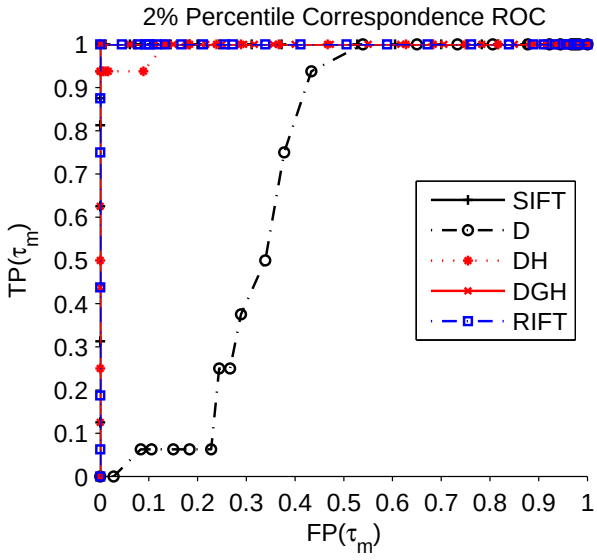


(a) Revolver

(b) Pistol (second reference)



(c) Apple iPod



(d) Binoculars

Figure 5.15: ROC curves when using percentile matches ($p = 2\%$) for correspondence set

descriptor has the broadest peak, closely followed by the DGH descriptor. The SIFT descriptor, though apparently with near-perfect ROC, only just reaches a peak of 1.0 before falling away. The density histogram, though apparently not as good in the ROC plot (Figure 5.15d), has the widest peak which would indicate it is more tolerant to detection-threshold-selection error.

Varying threshold quality gives us an alternative statistical visualization of the relative performance of the different 3D interest-point descriptors within this context.

5.5 Conclusions

Our results have shown that creation of the correspondence set using the distinction method of Lowe (2004) is not the best approach in the case of complex CT imagery containing a large number of artefacts. Better results are obtained if the correspondence set is determined by sorting the matches by Euclidean match distance and then taking a fixed percentage of the best matches.

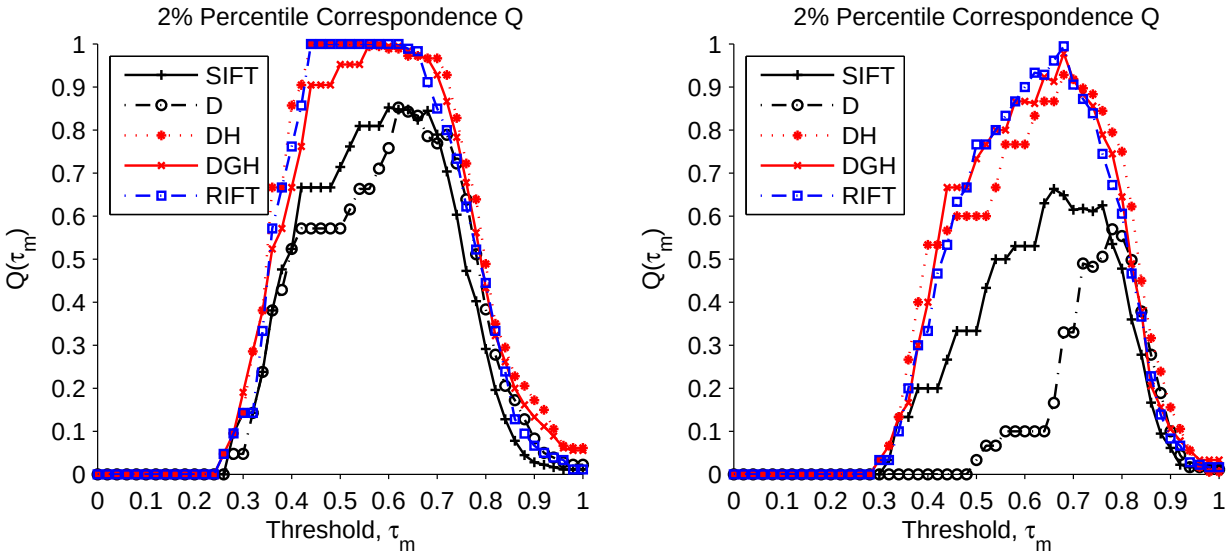
Detection of the revolver, pistol and binoculars was achieved with near-perfect results although this is more an indication that the number of scans containing the target items needs to be increased to correctly estimate margins of error in detection. Explicit cross-class recognition tests were not performed though some of the items were present in baggage items. For example, the iPod and binoculars appeared in the same baggage item so there was some degree of cross-testing in this case.

We have shown that an anisotropic-scanning system will affect the recognition results. The Browning pistol was scanned in orthogonal orientations and produced very different recognition results. Care thus needs to be taken when choosing a reference item or, as we have demonstrated, multiple reference volumes can be used to improve detection results. The use of multiple reference-object scans and methods of determining reference-scan quality is also left as an area for future work (when more data might be available).

By contrast to the complexity of the 3D-SIFT implementation, a simple histogram of density data in the local region of a point of interest provided very good comparative results.

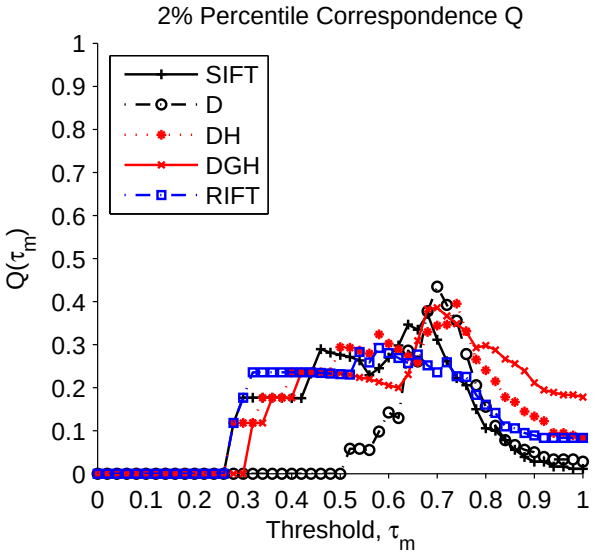
The 3D-RIFT descriptor produced good results using the distinction approach to produce the correspondence set and also performed well in the fixed-percentage approach. The 3D-RIFT descriptor is very concise: only 8 values are stored compared to 864 for 3D SIFT.

The 3D-SIFT descriptor can produce good results but it would appear that

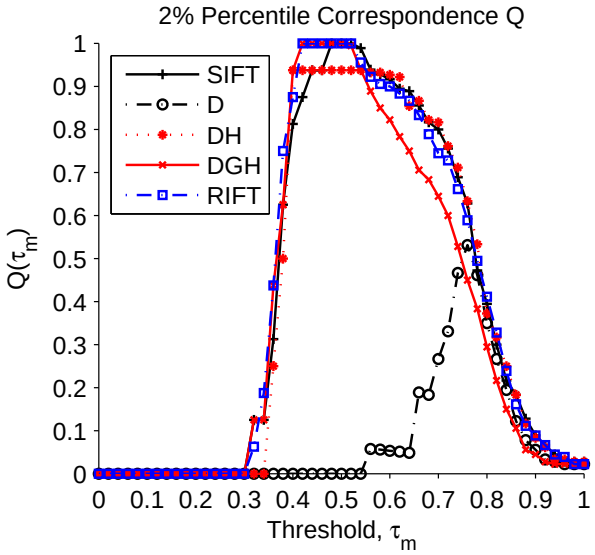


(a) Revolver

(b) Pistol (second reference)



(c) Apple iPod



(d) Binoculars

Figure 5.16: Threshold quality for percentile ($p = 2\%$) correspondence set

simpler descriptors (DH, 3D RIFT) produce better results with the advantage of reduced complexity. It would appear that the 3D-SIFT descriptor is not robust in the presence of a large amount of CT artefacts and this is understandable as the artefacts will greatly affect the density gradients upon which dominant orientation is decided and subsequent descriptor histograms are built.

Detection of the iPod was poor. The best result of 30% was achieved using the percentile method ($p = 2.0\%$). It is believed this is due to its lower density which is more easily corrupted by metal artefacts in the baggage item. It is also a fact that the iPod dimensions ($104mm \times 62mm \times 11mm$) ensure that most descriptors include areas outside the device in their formulation and, as such, are prone to adjacent baggage items influencing the descriptor.

Overall we have shown a comparison of differing 3D-point descriptors applied to the problem of object detection in complex 3D-CT-volumetric imagery. It has been shown that approaches based on simpler density information outperform more complex 3D extensions of common and established point descriptors adapted from 2D-image recognition (Lowe, 2004; Lazebnik et al., 2005).

The object-detection methodology so far described has shown that detection of known items can be successfully achieved under several assumptions. However, this method is impractical when considering a real airport scenario. At present we would require at least one reference scan for every potential target item and, given that there may be thousands of such items, there are issues regarding data storage and time taken to verify a match. This approach is also fallible if a previously unseen threat item (for example a new design of pistol) is presented. A more practical approach would be to establish the salient features that define a general class of objects, generalizing a training set or design to cover unseen instances of a class. and use these to determine the presence of a threat item within the baggage. We now seek to address the weakness of the current approach by exploring machine-learning techniques in the detection of object classes rather than specific objects.

Chapter 6

A codebook approach to object detection

We extend the work from the recognition of individual objects to object classes through creation of a visual word codebook and use of machine-learning approaches (van Gemert et al., 2010). To this end we aim to characterize a given class of object as a codeword within a codebook and use machine-learning techniques to identify an object as a particular class from its codeword description.

6.1 Introduction

Searching for specific known items has obvious limitations when considering the baggage-scanning environment. There are many shape variations within a class of threat items (handguns for instance in Figure 3.12) and each example would require its own reference scan(s) to be taken, stored and searched for in the baggage item. A better solution would be to learn the salient characteristics of a particular class of item and use those to search the baggage.

In this case we employ the bag of visual words approach (Sivic and Zisserman, 2003; Csurka et al., 2004) which seeks to reduce the many descriptors derived from a set of baggage items into a fixed length dictionary or codebook of visual words. Each codebook entry is akin to a word in a vocabulary: some words will be important in the description of the threat items. The descriptors from each baggage item are examined and assigned to a codebook entry. A histogram of codebook assignments can then be produced for the bag and this histogram now represents its visual word description. Given that the descriptors relate to points of interest in the baggage item, we can see that each entry in the codebook can be interpreted as a representation of whether a particular feature is present in the bag. For instance,

one codebook entry could be representative of a handgun trigger; should that entry appear in a baggage histogram, it could be indicative of a handgun being present.

Given a set of baggage items containing both positive and negative examples of a known class of threat, we anticipate that their codebook histograms can be used to determine whether a baggage item contains a threat item. This is achieved through supervised training of a SVM classifier using a training set of codebook histograms. Testing of the classifier can be performed using a set of unseen baggage items from which we can produce both true positive (correct detection of threat item) and false positive (incorrect classification of non-threat item as threat) metrics.

Forming the codebook from the descriptor set is achieved through clustering. The most common method is K-means clustering (MacQueen, 1967) which is an iterative approach that forms clusters of descriptors so that the sum total variance between each descriptor and its assigned cluster is minimized. Clustering identifies locations in descriptor space that are common to a cluster of contributing descriptors. Each cluster location then represents one visual word to be encoded in the codebook. The Euclidean distance between each descriptor and the cluster locations is used to assign the descriptor to the codebook histogram. The generation of the codebook is referred to as *vector quantization* (Sivic and Zisserman, 2003).

Allocation of a descriptor to the codebook has traditionally been performed by a hard assignment: choosing the codebook entry that is closest to the descriptor and incrementing the count of that codebook entry. However, recent work (van Gemert et al., 2010; Yang et al., 2007; Philbin et al., 2008) has shown that a soft-assignment methodology improves performance. Soft-assignment methodologies aim to address situations where the allocation of a particular descriptor to the codebook is ambiguous, overcoming the situation where a descriptor is only marginally closer to one codebook entry than another.

In this work we seek to assess the performance in detection of a class of items within our imagery whilst varying the codebook assignment methodology using state-of-the-art techniques (van Gemert et al., 2010) and varying the number of visual words employed in the codebook. Two classes of object are examined: handguns and bottles. We also seek to examine, given the results in Chapter 5, how the choice of descriptor affects performance.

6.2 Interest point locale and description

The location of interest points within the volumetric imagery is performed using the same parameters and methodology described in Chapter 4.

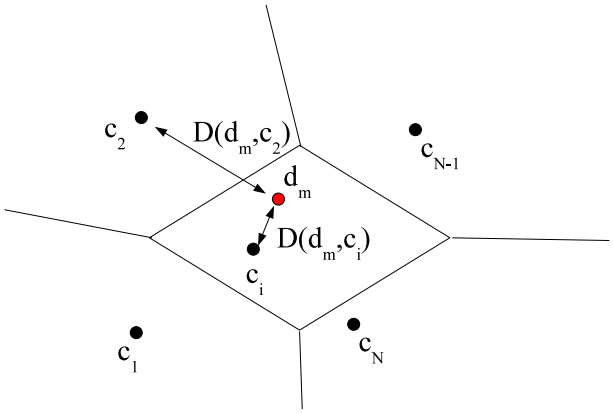


Figure 6.1: Codebook assignment in 2D

Following the results of Chapter 5 we examine the use of density histogram (DH), density-gradient histogram (DGH), rotation invariant feature transform (RIFT) and SIFT in this section of work. The density descriptor (Section 5.2.2) is not analyzed as its poor performance in the prior work did not warrant extension to this section.

Subsequent to interest-point location, each descriptor is generated using the methodologies and parametric settings given in Chapter 5.

6.3 Codebook formulation

Given a set of descriptors from scans of a series of baggage items, we use the K-means-clustering algorithm (MacQueen, 1967) to calculate a set of cluster centres for that group. We then apply vector quantization using the cluster centres to formulate the bag-of-words vector (Sivic and Zisserman, 2003) for each baggage item using three different codebook-assignment methods. The work of van Gemert et al. (2010) has shown that a soft assignment based on codeword uncertainty outperforms hard and kernel-based codebook assignment. We seek to replicate this work in our application to verify that these results hold true given the poor quality of our 3D imagery.

Given K clusters we will have a vocabulary comprising K codebook words in descriptor space. Assume that there are M descriptors contributing from the volume to the codebook histogram. If we consider the cluster centres, c_i $(i = 1 \ldots K)$, and descriptors, d_m $(m = 1 \ldots M)$ we can take the Euclidean distance between the descriptors and cluster centres, $D\left(d_m, c_i\right)$, as the measure of similarity, as shown in Figure 6.1. We build the codebook with words, w_i $(i = 1 \ldots K)$, using hard, kernel and uncertainty-assignment methods (van Gemert et al., 2010) as defined in the following sections.

6.3.1 Hard assignment

Hard assignment is the original codebook-assignment approach whereby every descriptor (d_m) is assigned to a cluster centre, c_i, minimizing the distance $D(d_m, c_i)$ over all c_i. Essentially the closest cluster centre to the descriptor is taken as the assignment. This can be formulated mathematically as follows:

$$CB_H(i) = \frac{1}{M} \sum_{m=1}^{M} \begin{cases} 1 & \text{if } w_n = \arg\min\left(D\left(d_m, c_i\right)\right) \\ 0 & \text{otherwise} \end{cases} \tag{6.1}$$

where K is the number of clusters and M is the number of descriptors in the volume. Note the normalization by M to ensure that the same histogram would be produced regardless of the number of contributing descriptors. This is the simplest form of assignment used in conventional codebook approaches (Sivic and Zisserman, 2003; Csurka et al., 2004).

6.3.2 Kernel assignment

A disadvantage of hard assignment is that it allows for no uncertainty in the formulation of the codeword assignment. It has been shown that this can degrade performance when compared to assignment methods that allow for some degree of fuzziness in the assignment (van Gemert et al., 2010). To overcome the problems associated with a hard-assignment methodology, a simple Gaussian kernel can be used to provide the assignment ambiguity in the codebook, assigning values as a function of distance from the descriptor to the cluster centre:

$$CB_K(i) = \frac{1}{M} \sum_{m=1}^{M} \exp\left[-\frac{1}{2}\left(\frac{D\left(d_m, c_i\right)}{\sigma}\right)^2\right]. \tag{6.2}$$

Here K is the number of clusters and M is the number of descriptors in the volume. The value of the smoothing parameter, σ, determines the degree of assignment fuzziness and hence the degree to which assignments to adjacent clusters are made. Again note the normalization by M to ensure that the same histogram would be produced regardless of the number of contributing descriptors.

Kernel assignment has a drawback: the kernel may assign a low value to the codebook for a descriptor even though it would appear to a human to be the most likely entry. Figure 6.2 shows such a situation where it seems obvious that descriptor d_m should have a strong contribution to the i^{th} codebook entry given that codeword c_i is closest. However, in this case the smoothing parameter, σ, is much smaller than the distance between the descriptor d_m and the codeword c_i, so that the derived

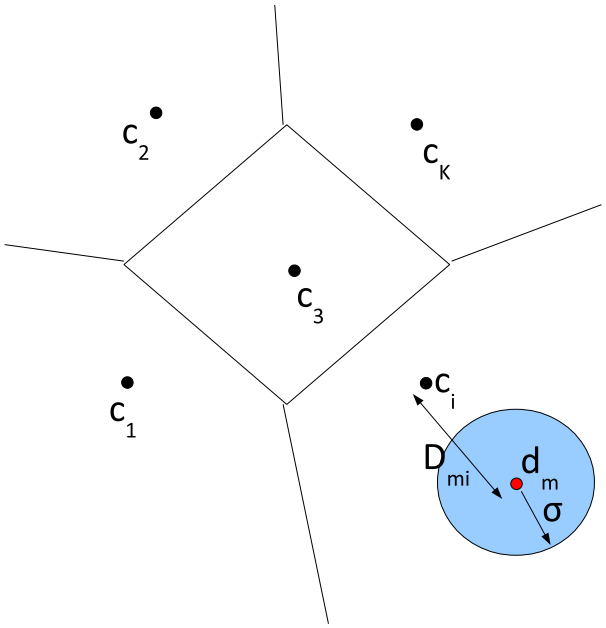


Figure 6.2: Kernel-assignment flaw

kernel value creates a small assignment to the codebook.

6.3.3 Uncertainty assignment

A normalization process can be employed to overcome the potential kernel-assignment flaw (Section 6.3.2) such that each descriptor contributes the same sum value to the codebook. This normalization ensures that each descriptor only contributes a sum total of 1.0 to the codebook and removes the low value (weak contributions) that can occur using kernel assignment (Section 6.3.2). The uncertainty assignment is given by the following:

$$CB_U(i) = \frac{1}{M} \sum_{m=1}^{M} \frac{\exp\left(-\frac{1}{2}\left(\frac{D(d_m,c_i)}{\sigma}\right)^2\right)}{\sum_{j=1}^{K} \exp\left(-\frac{1}{2}\left(\frac{D(d_m,c_j)}{\sigma}\right)^2\right)} \tag{6.3}$$

Here K is the number of clusters and M is the number of descriptors in the volume. The value of the smoothing parameter, σ, again determines the degree of assignment fuzziness and hence the degree to which assignments to adjacent clusters are made. Again note the normalization by M to ensure that the histogram is not biased by the number of descriptors obtained for a particular volume when it is compared to the histogram obtained for a different volume.

The work of van Gemert et al. (2010) found that this approach produced the highest true-positive rates in the task of scene classification using 2D imagery.

6.4 Detection methodology

An overview of descriptor generation is shown in Figure 6.4 where we see the separation of interest-point detection from descriptor generation which, in our comparison for this work, can be performed in a number of different ways (as described in Section 5.2). Interest-point locations for an input volume are generated using the Difference-of-Gaussian methodology described in Section 4.2 and the descriptors for each volume are generated using the range of methodologies described in Section 5.2.

Following descriptor generation we proceed with vector quantization as described in Section 6.3. We employ a ten-fold cross-validation method to derive training and test sets.

An overview of the codebook approach is shown in Figure 6.5. We start with a set of training volumes from which a training set of descriptors are generated. These descriptors are input to the K-means algorithm to derive a set of cluster centres. The K-means algorithm is initialized using the algorithm derived by Arthur and Vassilvitskii (2007) which has been shown to improve the convergence speed over a random initialization of the cluster centres. The K-means-clustering algorithm seeks to minimize the sum square distance from each cluster centre to the data points being processed: the cluster compactness. However, the algorithm is prone to sub-optimal solutions when the initial cluster centres are randomly assigned. To overcome this problem the algorithm is executed a number of times (10) and the result with the minimal cluster compactness chosen. The resulting cluster centres are the codebook words. The training descriptors then undergo vector quantization using the chosen assignment method and appropriate settings for codebook size (K) and smoothing parameter (σ) as described in Section 6.3.

For classification we use a support vector machine (SVM). The SVM is presented with a training set of data containing two classes. A simple example is shown in Figure 6.3a where we can see two classes of objects. The SVM attempts to solve this classification problem by moving the data to a higher dimension so that a hyperplane can be fitted that separates the two classes and maximizes the margin between them (Figure 6.3b). In some instances a soft margin can be used that allows for misclassification, as shown in Figure 6.3c. The degree to which the soft margin applies is set in the SVM by a parameter called cost, C. This parameter allocates a weighting to misclassified examples when the SVM is training with the aim of minimizing the total misclassification error. Another aspect of the SVM is the hyperplane. In its simplest form a linear hyperplane is used. However, better classification results can be achieved using nonlinear hyperplanes. For our work we

used a SVM with a Gaussian-radial-basis function as its kernel, with parameter γ setting the Gaussian response.

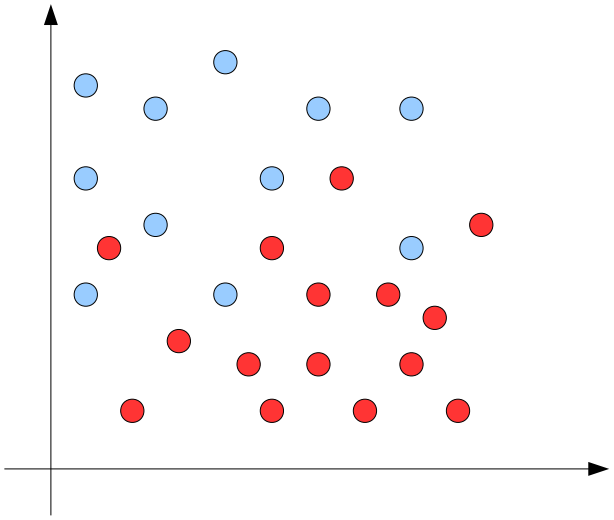
The training codebook vectors are used as input to a SVM (Cortes and Vapnik, 1995; Bishop, 2006) in order to generate a supervised-learning classifier. The classifier is given the training data from which it configures itself for the classification task in hand. A set of unseen test volumes are then processed to generate descriptors that are vector quantized in the same manner as the training descriptors. The test vectors are then input to the SVM which classifies the volumetric images according to the presence of a threat item or not. These classification decisions are noted and from this we can determine true-positive (TP) and false-positive (FP) detection rates for threat items in the CT-volumetric data as we know the contents of the test volumes (see Appendix C).

We vary the number cluster centres, K, the smoothing parameter, σ, as well as the classifier configuration in order to investigate the performance of the detection system for each of the descriptors being evaluated. The SVM is implemented using a Radial Basis Function (RBF) kernel that requires setting of two parameters: C and γ. The optimal values of (C, γ) are derived using a grid-search approach. For each setting of (C, γ) in the grid, a ten-fold cross-validation is performed using the training vector set. The number of misclassifications is noted and the process repeated for new settings of (C, γ). The values of (C, γ) that result in the minimum number of misclassifications (both target object and clutter) are chosen as the configuration for final training and testing.

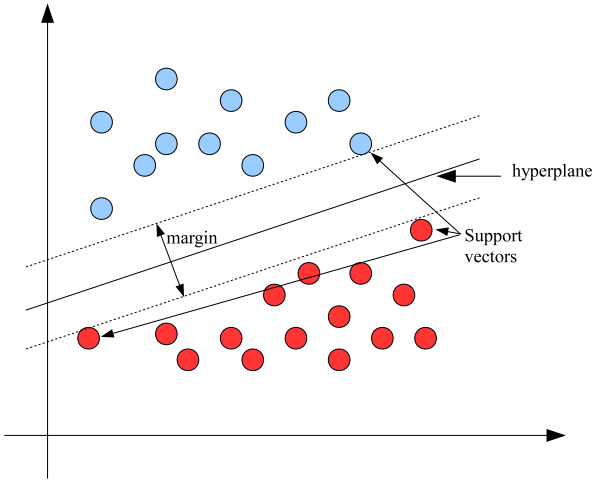
6.4.1 Data sets

Our training and testing datasets were derived from CT scans of baggage items containing clutter with or without a threat item. Two threat items were considered. The first group of threat items consisted of handguns (pistols and revolvers). The second group of threat items were bottles of varying shapes and sizes that contained liquids of various densities and in varying quantity.

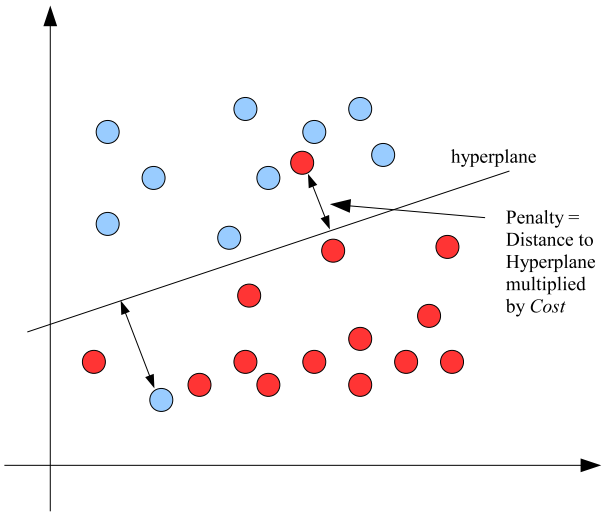
For each threat group we derive two distinct datasets within this work. The first dataset is constructed by cropping the volumetric data to isolate the threat items. A margin of $5cm$ around the threat item was included and these cropped sub-volumes provided the threat dataset. Examples of cropped handguns are shown in Figure 6.6a. Examples of cropped bottles are shown in Figure 6.6b. Baggage items which did not contain a threat were subdivided into volumes similar in size to the threat item sub-volumes. These sub-volumes provide the clear dataset. Examples of non-threat sub-volumes for the handgun group are shown in Figure 6.7a and examples



(a) Input classification data



(b) Linear hyperplane at higher dimension separates classes



(c) Allowing misclassification: Cost

Figure 6.3: Example SVM classification task

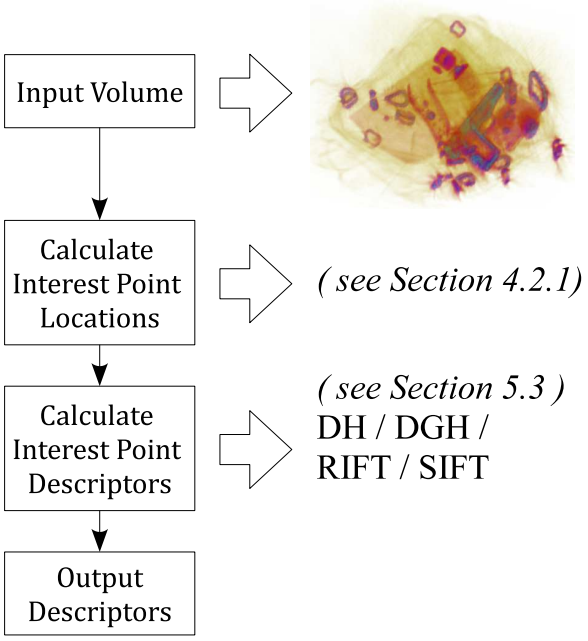


Figure 6.4: Descriptor generation

Item	Quantity
Threat	284
Clear	971

(a) Handgun dataset

Item	Quantity
Threat	526
Clear	1178

(b) Bottle dataset

Table 6.1: Sub-volume data sets

used for the bottle group are shown in Figure 6.7b. Table 6.1a shows the number of items in the sub-volume category for the handgun threat item and Table 6.1b the number of sub-volumes when bottles were used.

Given the amount of baggage data available for this study it was not possible to acquire sufficient handgun-only scans, bottle-only scans and clutter-only scans. Consequently handguns are considered to be clutter when the bottle is marked as the threat and vice versa.

6.5 Results using handgun sub-volumes

We will now present the results for the detection of handguns in the baggage items from the data generated in Section 6.3.

For each baggage item given Table 6.1a sets of descriptors were calculated (DH, DGH, RIFT and SIFT). For each descriptor we then performed the following ex-

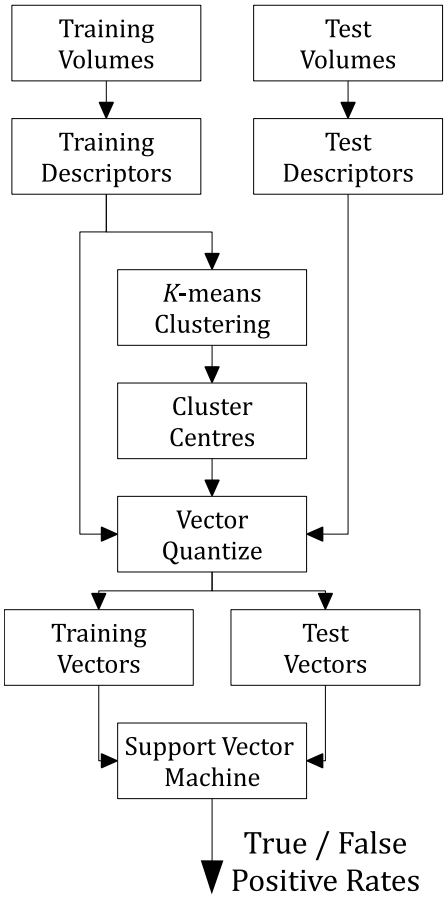
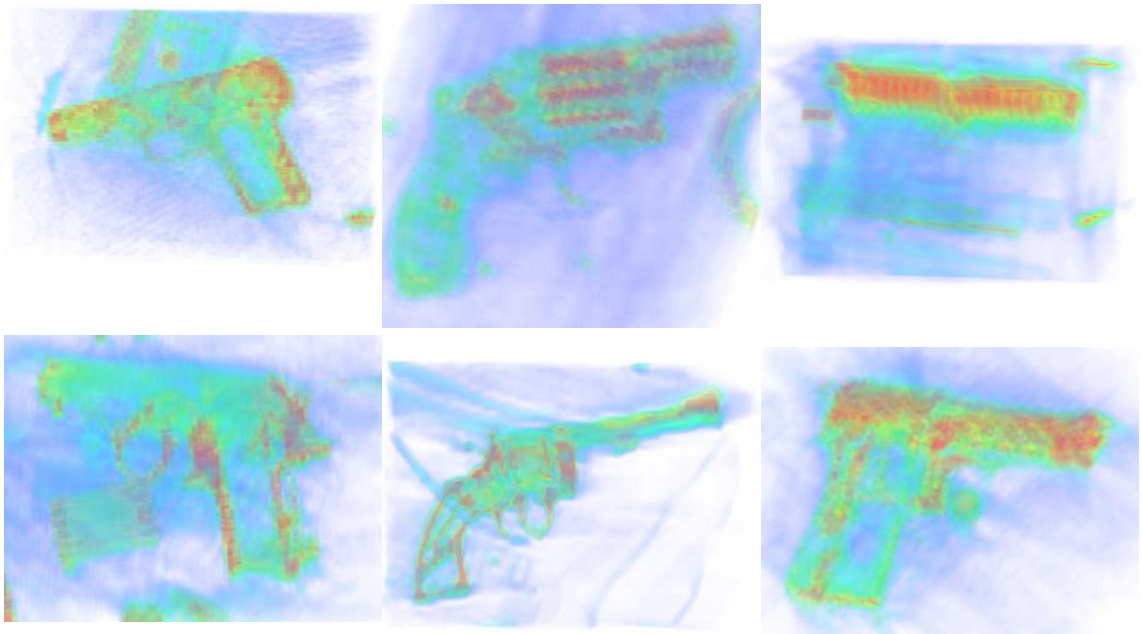
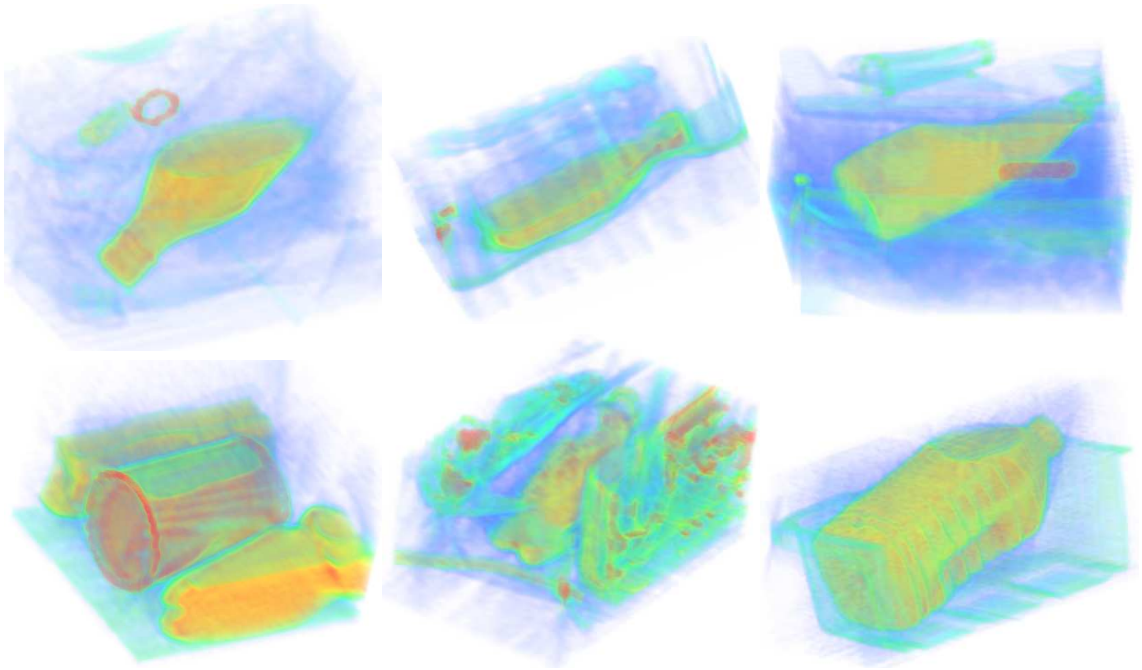


Figure 6.5: Bag of words approach

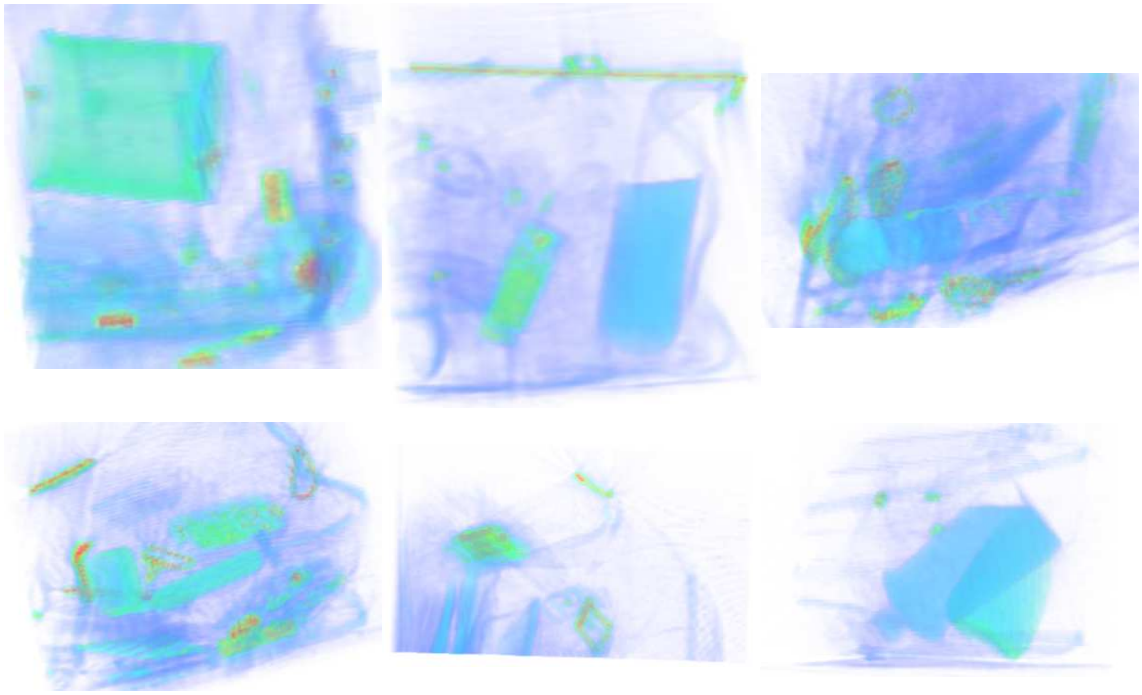


(a) Handgun threat items

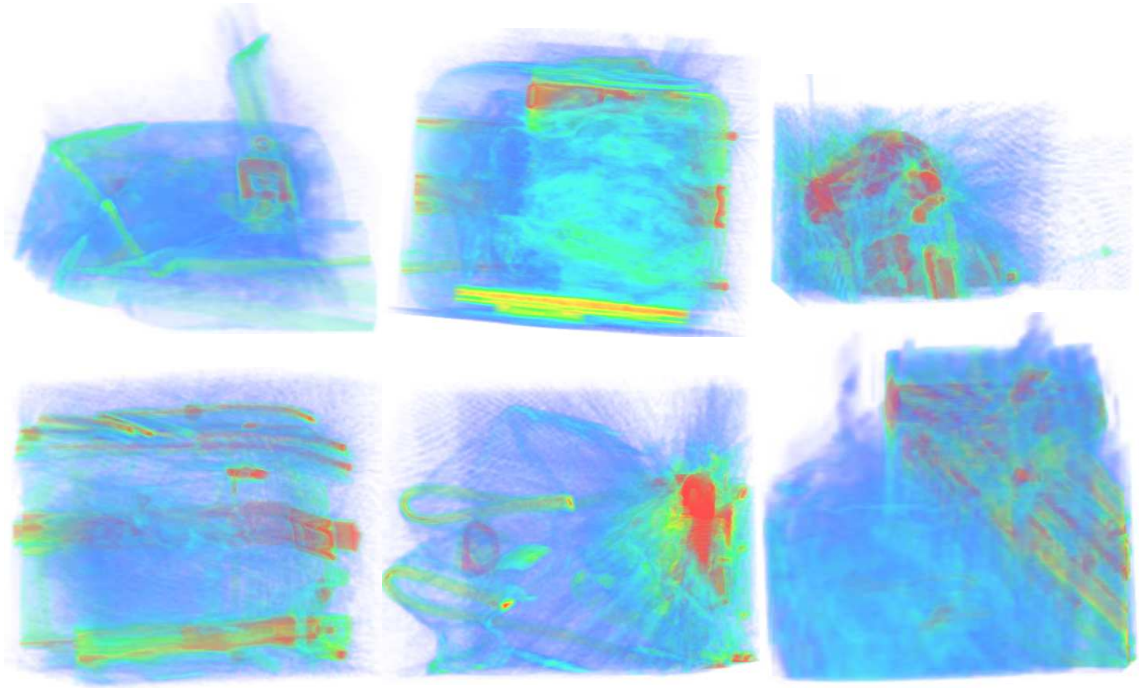


(b) Bottle threat items

Figure 6.6: Example threat sub-volumes



(a) Handgun group: no threat sub-volumes



(b) Bottle group: no threat sub-volumes

Figure 6.7: Example clutter sub-volumes

periment. Codebooks were generated using hard-assignment, kernel-assignment and uncertainty-assignment methods, with a range of parameter values (K, σ), as described in Section 6.3. For each resulting codebook we performed a ten-fold cross-validation procedure, recording the mean and standard deviation for both true-positive and false-positive rates, as described in Section 6.4. The standard deviation is recorded in tabular results in the normal manner or as an error bar in graphical results.

6.5.1 Parameter setting for kernel and uncertainty assignment

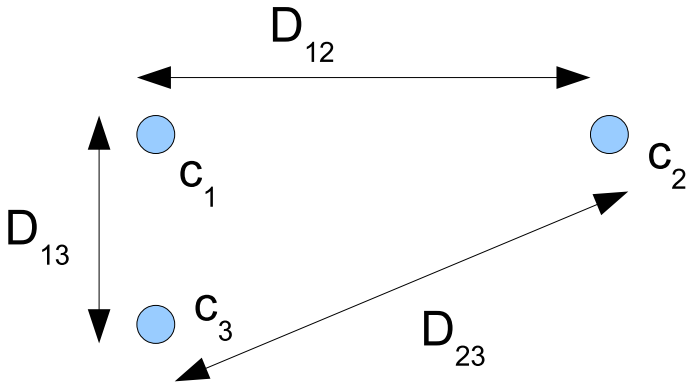
Both kernel and uncertainty-assignment methods use the smoothing parameter, σ, to define the influence on neighbouring clusters (Equations (6.2) and (6.3)). In these formulations, the value of σ defines the distance over which the assignment has an effect. The choice of a suitable value for σ will depend on the cluster spacings for a given set of data and value of K. We examined the cluster distributions for the handgun sub-volume data set.

The methodology used to evaluate the clustering will now be given. A simple example is shown in Figure 6.8 for the case of $K = 3$. In Figure 6.8a we see three cluster centres (c_1, c_2, c_3) separated by distances D_{12}, D_{13} and D_{23}. A matrix showing the distances between each cluster is shown in Figure 6.8b. We now sort the elements of each row in this matrix in ascending order so that we can see the distance from each cluster centre to the n^{th} furthest away ($1 \leq n \leq K$). This is shown in Figure 6.8c where we can also see that the mean of each column, m_n, is calculated to give us the mean distance to the n^{th} furthest cluster. The second column of this matrix will yield the mean distance to the nearest adjacent cluster centre for a given K-means-clustering operation.

We now consider the case of 1024 clusters for each of the descriptors being investigated. Following K-means clustering, there will be 1024 cluster centres defined in descriptor space. We calculate the distance from each cluster to its neighbours then sort by ascending distance so that we can see how close each cluster is to its neighbours using the methodology described in Figure 6.8.

Figure 6.9a shows the mean distance to the n^{th}-sorted cluster where we can see the nearest adjacent clusters are a Euclidean distance of less that 0.2.

A closer investigation on the nearest adjacent cluster distances (i.e. column 2 from Figure 6.8c) can be made to examine their distribution. Figure 6.9b shows a histogram of the nearest adjacent cluster distances using the handgun sub-volume data-set with 1024 clusters where we can see peaks in adjacent cluster distance in



(a) Three example cluster centres

From \ To	c_1	c_2	c_3
c_1	0	D_{12}	D_{13}
c_2	D_{12}	0	D_{23}
c_3	D_{13}	D_{23}	0

(b) Matrix recording the distance between each cluster

Cluster \ n	1	2	3
c_1	0	D_{13}	D_{12}
c_2	0	D_{12}	D_{23}
c_3	0	D_{13}	D_{23}
mean	m_1	m_2	m_3

(c) Sorting each row in ascending distance then calculating mean of each column gives mean distance to n^{th} furthest cluster

Figure 6.8: Cluster distance and sorting

the region 0.02 to 0.06. The SIFT, DH and DGH descriptors all have peaks in the region 0.05 to 0.06 whereas the RIFT descriptor peak is nearer to 0.025, though its distribution is quite broad.

The value chosen for σ determines the distance over which the assignment has an effect (Equation (6.2), Equation (6.3)). The location of peaks in adjacent cluster distance (Figure 6.9b) occur for distances in the region 0.02 to 0.06. These values indicate the value of σ that should be used for the kernel and uncertainty-assignment methodologies when using 1024 clusters.

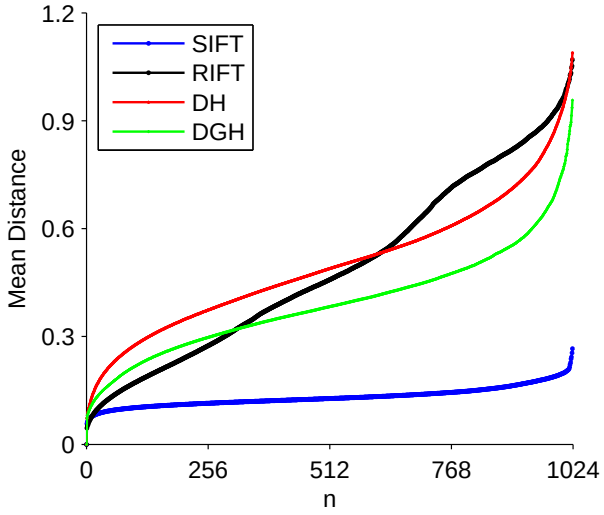
The number of clusters will be varied and so we need to consider how these measures vary with the number of clusters being used.

Figure 6.10 shows the same plots when 128 clusters are used. Figure 6.10b shows the SIFT descriptor peak at a distance of 0.05 but the DH and DGH distributions have all moved from ~ 0.06 to ~ 0.09 when compared to Figure 6.9b suggesting a relationship between adjacent cluster distance and number of clusters used. This makes sense as the distance between cluster centres will tend to reduce as the number of clusters is increased. Note that adjacent cluster histogram (Figure 6.10b) is more coarse in nature because there are fewer cluster centres ($K = 128$) than for Figure 6.9b ($K = 1024$).

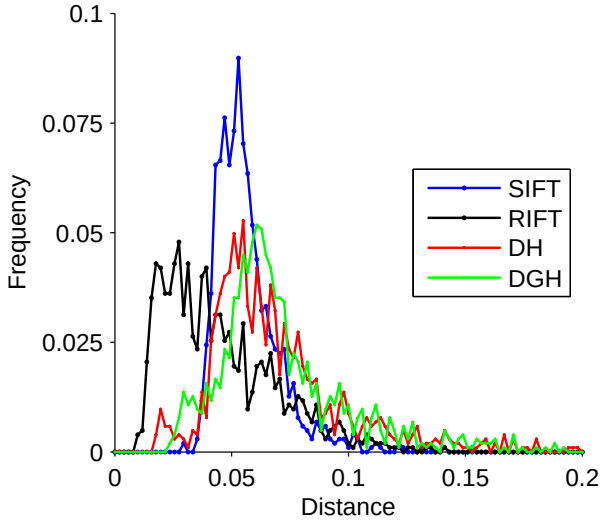
In our work we use a range of values for σ, for each setting of K, as it is not clear from the evaluation of cluster distances or from prior work in this area (van Gemert et al., 2010; Philbin et al., 2008) which value would result in the best performance. The choice of σ values is, however, guided by the location of peaks in Figure 6.9b and Figure 6.10.

6.5.2 Hard assignment

We use hard assignment (Section 6.3.1) whilst varying the number of clusters. The number of clusters is given by $K = 2^n$ where $n = \{6, 7, 8, 9, 10, 11\}$. The mean and standard deviation for the true-positive and false-positive detection results are taken from a ten-fold cross-validation using the handgun sub-volume volumetric data for each of the four descriptors being investigated. The results are shown in Figure 6.11. Figure 6.11a shows the true-positive detection results as K is varied for each of the descriptors. It can be seen from this that there is a distinct performance difference between the SIFT and RIFT descriptors when compared to the DH and DGH descriptors. We obtain detection rates of $\sim 80\%$ for SIFT and RIFT but DH and DGH both exceed 90% with the highest performance being 97.2% for DGH with $K = 2048$ and 96.1% for DH with $K = 256$. The best performance for SIFT is 83.0% for $K = 256$ and for RIFT it is 83.0% for $K = 512$. However, this is not the

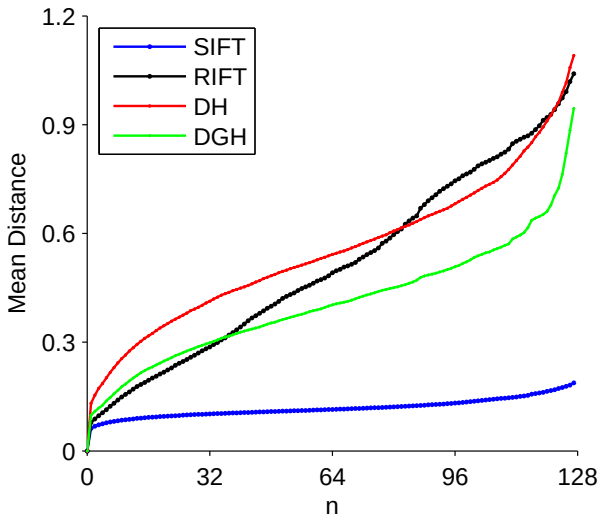


(a) Mean-sorted distance to n^{th} cluster

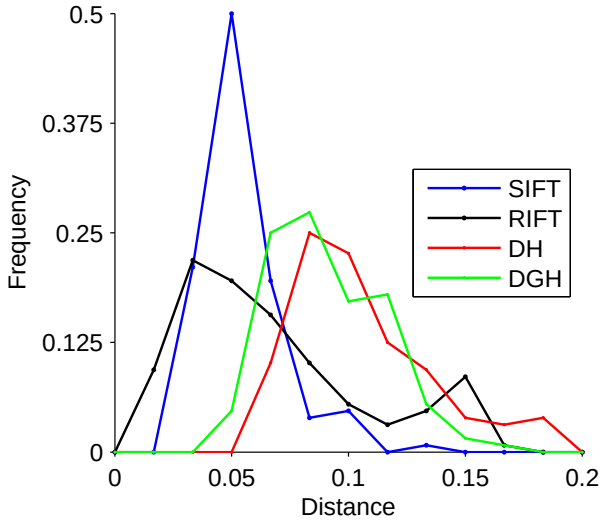


(b) Nearest adjacent-cluster distance histogram

Figure 6.9: Adjacent-cluster measures: K=1024

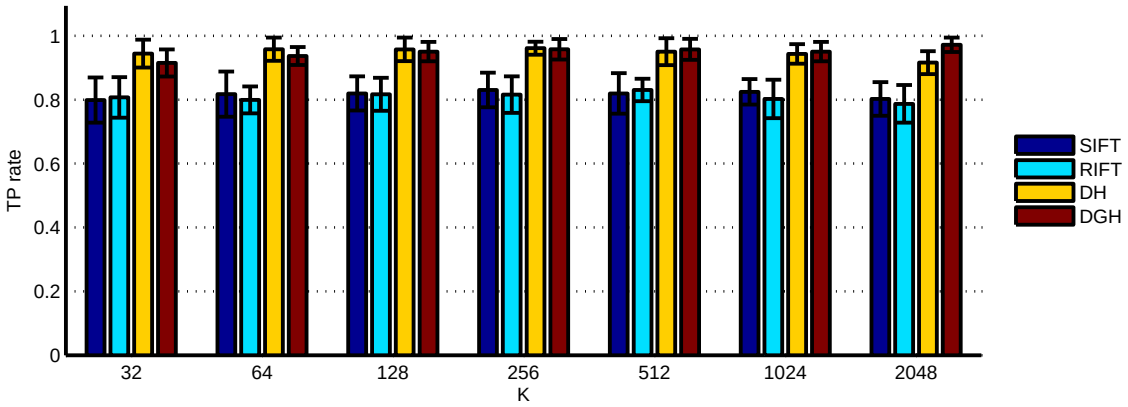


(a) Mean-sorted distance to n^{th} cluster

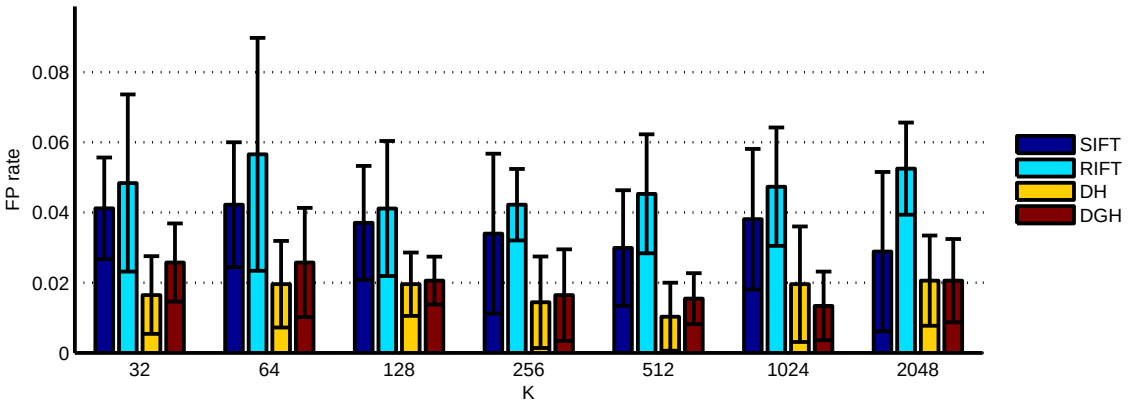


(b) Nearest adjacent-cluster distance histogram

Figure 6.10: Adjacent-cluster measures: K=128



(a) True-positive performance



(b) False-positive performance

Figure 6.11: Handgun sub-volume results using hard assignment

complete picture as we need to consider the false-positive performance.

False-positive results are shown in Figure 6.11b where we can again see a difference between SIFT/RIFT and DH/DGH. The lowest false-positive rate is obtained for DH with a rate of 1.0% for $K = 512$. The lowest false-positive rate for the SIFT descriptor is 3.0%, again with $K = 512$, and for RIFT it is 4.1% for $K = 128$. The density-gradient histogram (DGH) descriptor has its lowest false-positive performance for $K = 1024$ with a value of 1.3%.

In the transport-screening environment a high true-positive rate is desired together with a low false-positive rate (Shanks and Bradley, 2004). We summarize the performance using the settings (K, σ) that maximize the true-positive rate for each descriptor and these are given in Table 6.2.

Descriptor	K	TP rate (%)	FP rate (%)
SIFT	256	83.0 ± 5.4	3.4 ± 2.3
RIFT	512	83.0 ± 3.5	4.5 ± 1.7
DH	256	96.1 ± 2.0	1.4 ± 1.3
DGH	2048	97.2 ± 2.2	2.1 ± 1.2

Table 6.2: Handgun sub-volume best detection rates for each descriptor using hard assignment

6.5.3 Kernel assignment

With kernel assignment (Section 6.3.2) we vary both the number of clusters (K) as before as well as the smoothing parameter, σ. Based on the analysis in Section 6.5.1 we choose $\sigma = \{0.02, 0.04, 0.08, 0.16\}$.

Figure 6.12 shows results for the SIFT descriptor for each setting of the smoothing parameter and number of clusters. Here we see that for $\sigma = 0.02$ there is no classification result. This is explained by the nearest adjacent cluster distance histogram plot in Figures 6.9b and 6.10b where the SIFT descriptor has little, if any, adjacent clusters below a distance of 0.04. Consequently the kernel assignment in this case creates codebooks in which a large number of vector elements are 0.0 and the SVM fails to train adequately. When the smoothing parameter is more compatible with the distribution of clusters ($\sigma = \{0.04, 0.08, 0.16\}$) we see detection performance improve. It would appear that the best detection rate occurs for $K = 2048$, $\sigma = 0.08$ where there is a true-positive rate of 85.8% and a false-positive rate of 3.3%. Note how, for $\sigma = 0.04$, the true-positive rate improves as the number of clusters, K, increases. This is in line with earlier studies concerning two dimensional image descriptors and a codebook-based approach (van Gemert et al., 2010; Philbin et al., 2008). As the number of clusters is increased the distance between cluster centres will reduce with the result that a smaller setting of σ will have an increasing influence. This is seen most noticeably for $\sigma = 0.04$ (TP rate rising from 55% for $K = 32$ to 80.5% for $K = 1024$) but can also be seen to a lesser extent for $\sigma = 0.08$ (TP rate rising from 79% for $K = 32$ to 85.8% for $K = 2048$) and $\sigma = 0.16$ (TP rate rising from 80% for $K = 32$ to 84% for $K = 1024$). Note also that the best detection rate (85.8%) is slightly higher than the best hard assignment result (83.0%), this result is unlikely to be significant (as measured using a t-test, for example) as the difference lies well within the margin of error measured from the ten-fold cross-validation.

Figure 6.13 shows results for the RIFT descriptor for each setting of the smoothing parameter and number of clusters. Here we see detection for all values of the

Descriptor	K	σ	TP rate (%)	FP rate (%)
SIFT	2048	0.08	85.8 ± 4.3	3.3 ± 1.8
RIFT	1024	0.02	86.9 ± 5.4	4.7 ± 2.0
DH	1024	0.04	97.3 ± 3.4	1.8 ± 1.7
DGH	2048	0.04	96.8 ± 2.6	1.4 ± 1.3

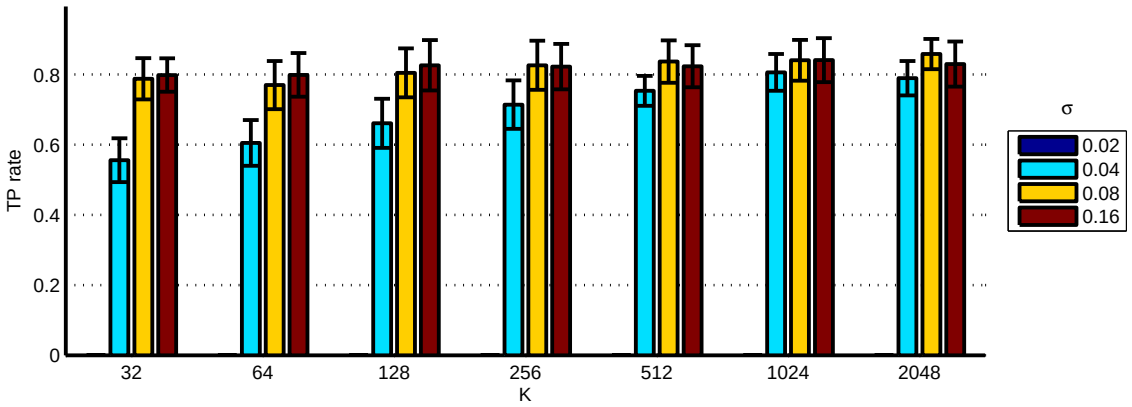
Table 6.3: Best detection rate for each descriptor using kernel assignment with SVM classifier

smoothing parameter. This reinforces the assessment given in Section 6.5.1 where it was observed that the RIFT-adjacent descriptor distances covered a broader range, starting at a lower value than SIFT. If we consider $\sigma = 0.02$ we again see a rise in detection rate as the number of clusters, K, increases and consequently the mean-cluster separation reduces starting with 64.8% for $K = 32$ rising to 86.9% for $K = 1024$.

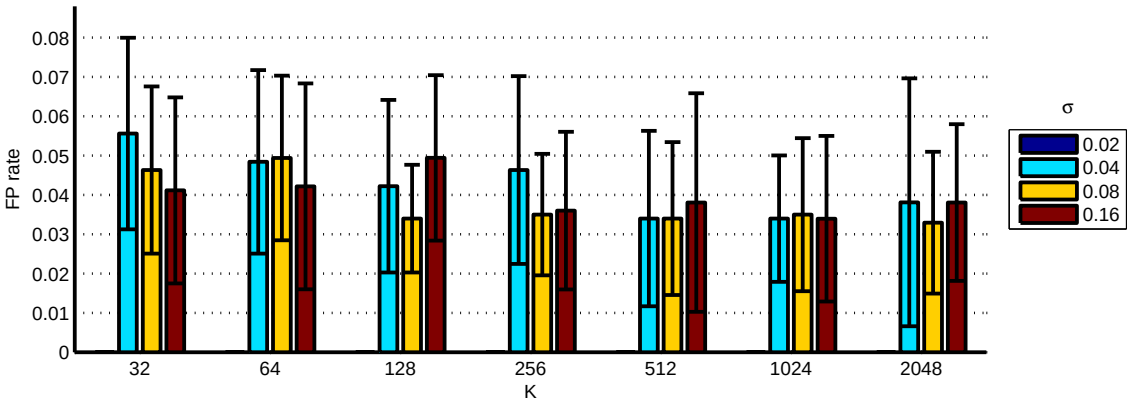
Figure 6.14 shows results for the density-histogram descriptor for each setting of the smoothing parameter and number of clusters. We can again see that setting $\sigma = 0.02$ is not appropriate but note that the performance improves as the number of clusters increases, as before. The same pattern is shown for $\sigma = 0.04$ until $K = 128$. The highest detection rate is 97.3% for ($K = 1024$, $\sigma = 0.04$), for which the false-positive rate is 1.8%.

Figure 6.15 shows results for the density-gradient histogram descriptor for each setting of the smoothing parameter and number of clusters. Again setting $\sigma = 0.02$ yield poor performance. The highest detection rate for this descriptor is 96.8% when ($K = 2048$, $\sigma = 0.04$) at which point the false-positive rate is 1.4%.

The best detection results for each descriptor are summarized in Table 6.3. From this we can make a number of observations. Firstly, given the measured measurement error, we can say that DH and DGH outperform SIFT and RIFT. We could argue that SIFT outperforms RIFT, based on the mean results for both true positive and false positive measures. Although the values of the smoothing parameter, σ, are quite coarse we can see that the value used for DH and DGH agree with the observations in Section 6.5.1. Likewise the setting for RIFT is in line with its lower distance histogram result (Figure 6.9b). With SIFT we see a higher setting of $\sigma = 0.08$ with similar results being obtained for $\sigma = 0.16$. These settings are above the peak in the distance histogram (Figure 6.9b) which would suggest that the SIFT clusters are less distinct as visual words and require more spread in the assignment.

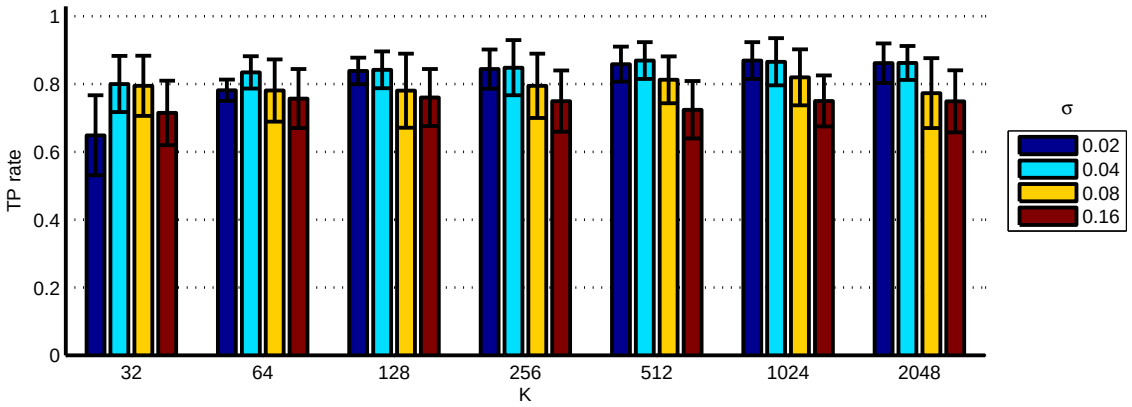


(a) True-positive performance

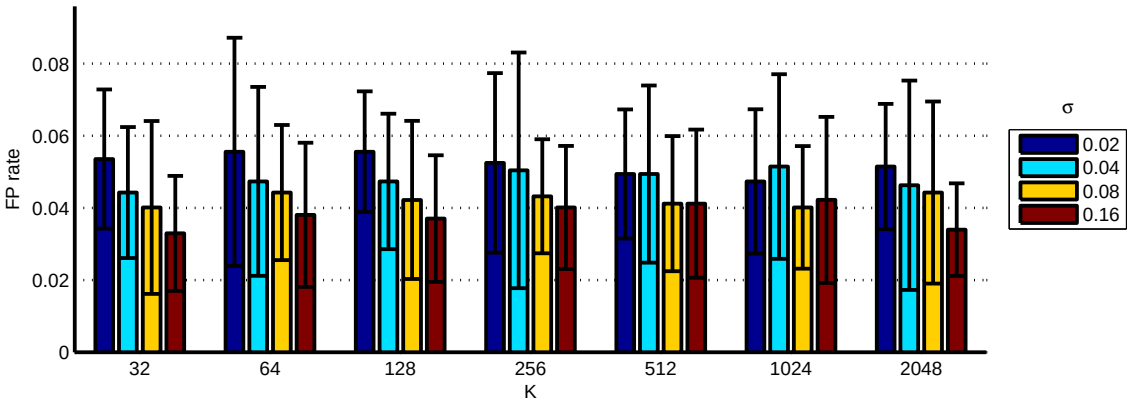


(b) False-positive performance

Figure 6.12: Handgun sub-volume results using kernel assignment, SVM classification for SIFT descriptor

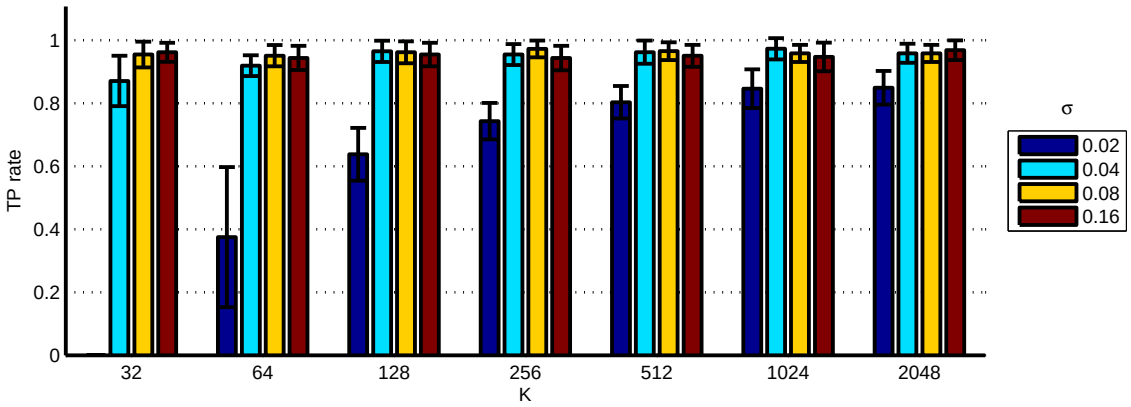


(a) True-positive performance

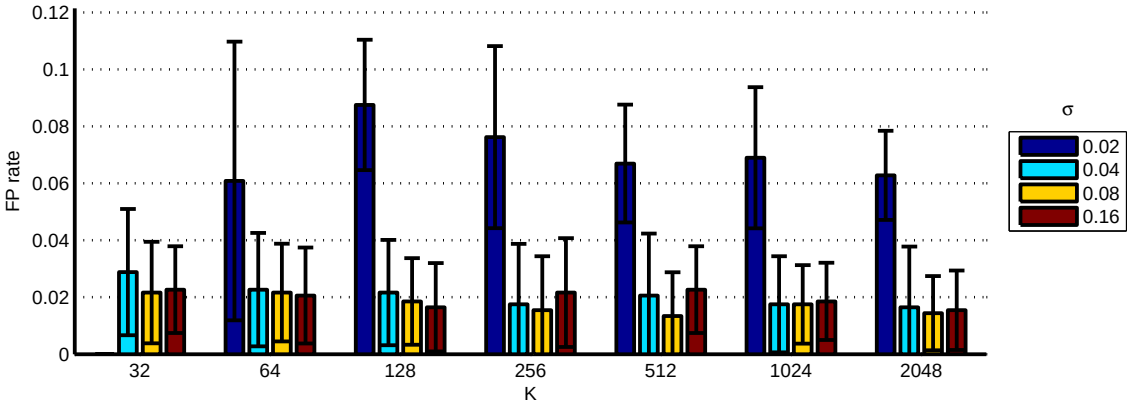


(b) False-positive performance

Figure 6.13: Handgun sub-volume results using kernel assignment, SVM classification for RIFT descriptor

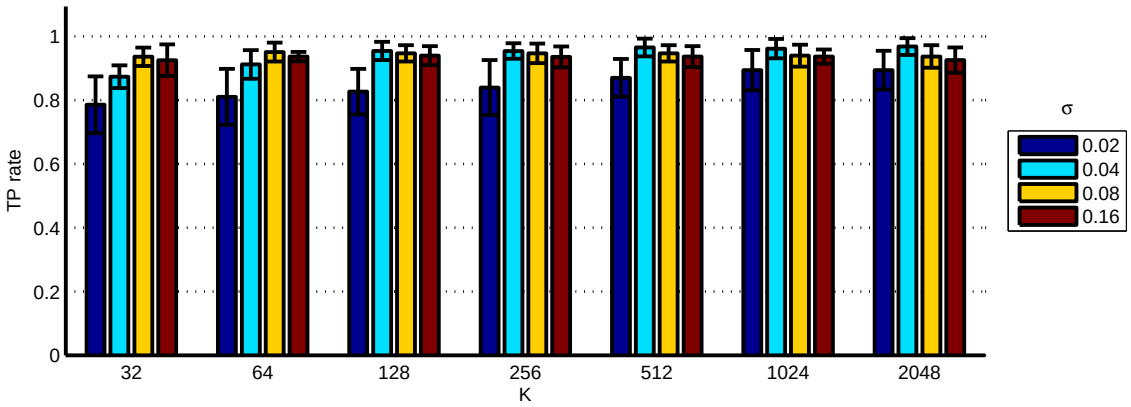


(a) True-positive performance

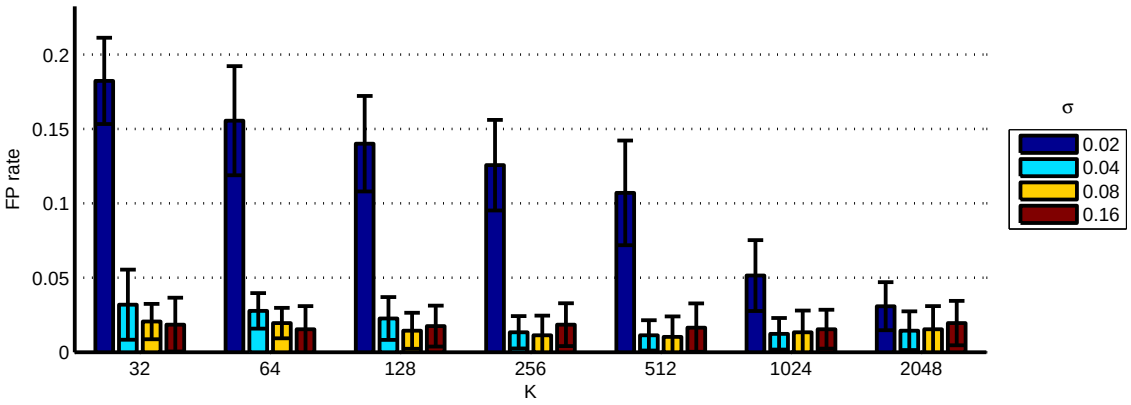


(b) False-positive performance

Figure 6.14: Handgun sub-volume results using kernel assignment, SVM classification for density-histogram descriptor



(a) True-positive performance



(b) False-positive performance

Figure 6.15: Handgun sub-volume results using kernel assignment, SVM classification for density-gradient histogram descriptor

6.5.4 Uncertainty assignment

For uncertainty assignment (Section 6.3.3) we initially follow the approach of kernel assignment and choose:

$$\sigma = \{0.02,\ 0.04,\ 0.08,\ 0.16\}$$

with:

$$K = \{32,\ 64,\ 128,\ 256,\ 512,\ 1024,\ 2048\}$$

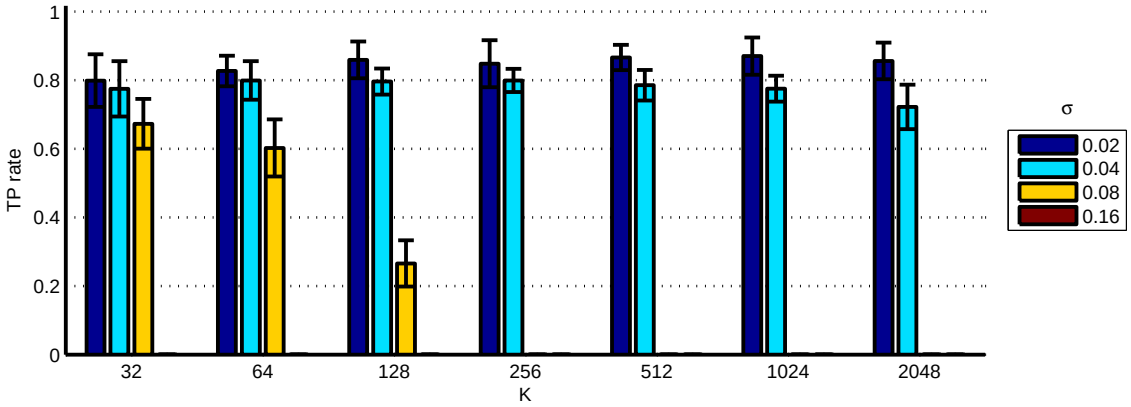
.

Results for the SIFT descriptor as shown in Figure 6.16 where we see a peak detection rate of 87.0% (Figure 6.16a) for ($K = 1024, \sigma = 0.02$) with a corresponding false-positive rate of 3.8% (Figure 6.16b). Both the true-positive and false-positive plots are interesting as they show poor performance for $\sigma = \{0.08,\ 0.16\}$, settings that produced the best results with the kernel-assignment method. No detection result is produced for $\sigma = 0.16$ for any value of K and results for $\sigma = 0.08$ are quite poor for $K = 32$ before tailing off to zero for $K = 256$. Given that the best performance was achieved using $\sigma = 0.02$ it was decided to extend the range of the smoothing parameter to lower values. Figure 6.17 shows the results for $\sigma = \{0.005,\ 0.01,\ 0.02,\ 0.04,\ 0.08,\ 0.16\}$ where we can now see reasonable performance for $\sigma = 0.01$, though not quite as good as when $\sigma = 0.02$. Poor detection rates are observed for $\sigma = 0.005$.

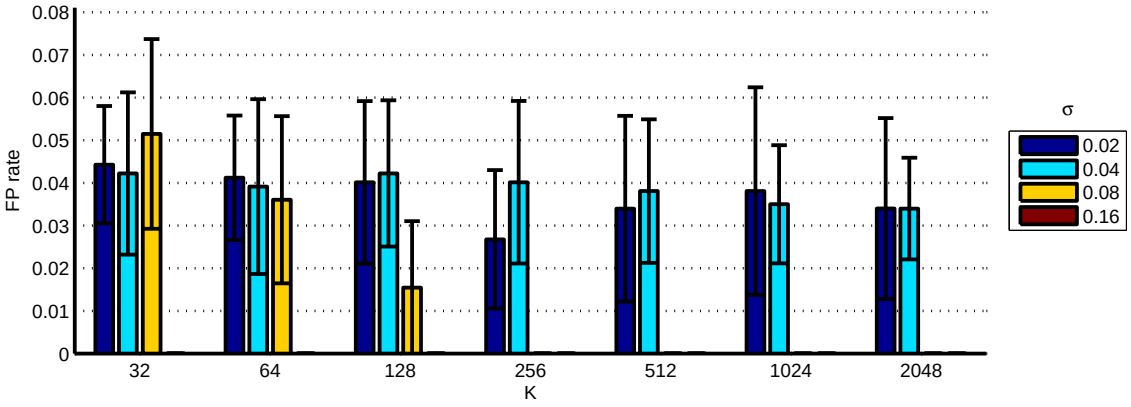
For the RIFT descriptor we again observed reduced performance for $\sigma = \{0.08,\ 0.16\}$ so extended the smoothing parameter to $\sigma = \{0.005, 0.01, 0.02, 0.04, 0.08, 0.16\}$ as for SIFT. The results for the RIFT descriptor can be seen in Figure 6.18 where we can see peak detection of 87.3% with a false-positive rate of 5.1% for ($K = 2048$, $\sigma = 0.01$). We can see from Figure 6.18a that performance for $\sigma = 0.005$ ranges from 74.8% when 32 clusters are used up to 81.9% for $K = 1024$.

Investigation using the density-histogram descriptor yielded the results shown in Figure 6.19. Peak detection is 97.2% for ($K = 2048$, $\sigma = 0.02$) with a false-positive rate of 1.6%. A high detection rate can be seen for all values of K although the value of the smoothing parameter needs to be accurately set at the extreme values of K. We can see the detection rate improve for $\sigma = 0.005$ as K increases and similarly we can see the performance for $\sigma = 0.16$ decline for values of K above 1024.

For the density-gradient histogram we see a peak detection rate of 97.2% with false-positive rate of 2.1% when ($K = 512$, $\sigma = 0.04$) as shown in Figure 6.20. The fall off in performance when $\sigma = 0.16$ is more noticeable (Figure 6.20a) when



(a) True-positive performance



(b) False-positive performance

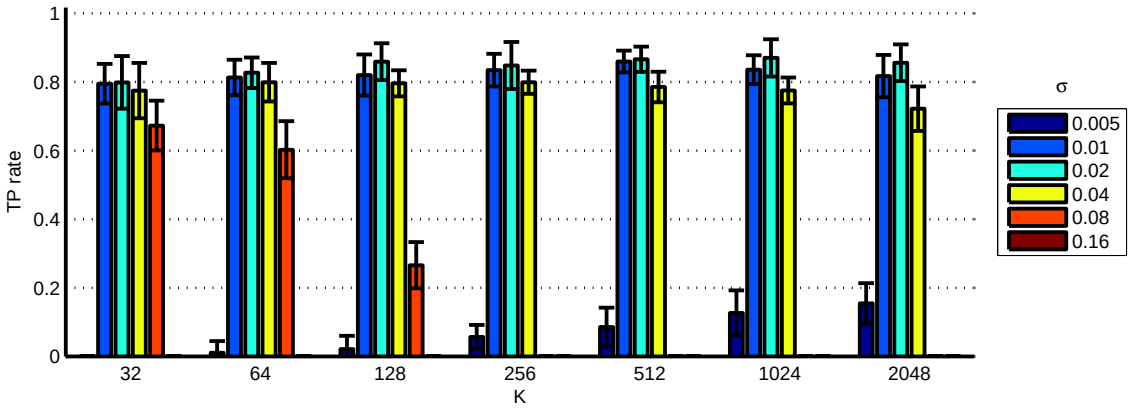
Figure 6.16: Handgun sub-volume results using uncertainty assignment, SVM classification for SIFT descriptor

compared to the density histogram (Figure 6.19a).

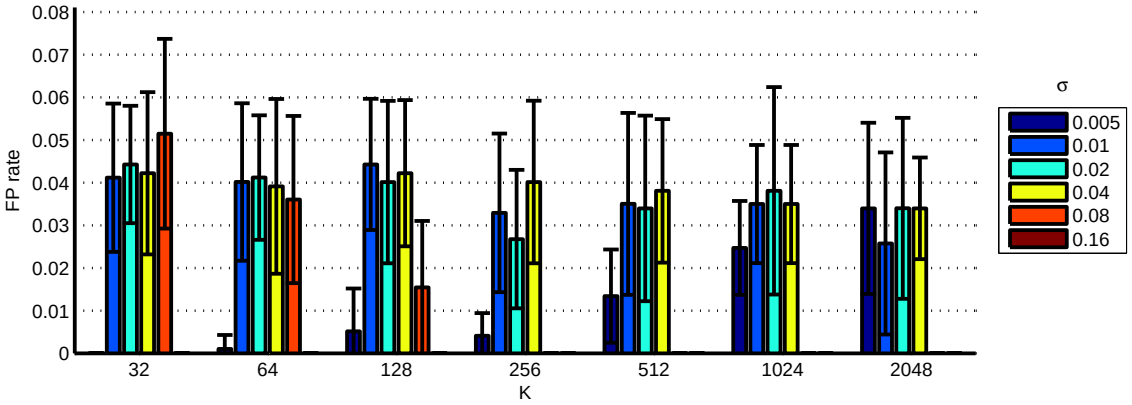
Table 6.4 shows a summary of best detection rates for each descriptor and the associated bag-of-words setting. We can see from this that both density histogram and density-gradient histogram have the highest detection rate (97.2%) with low false-positive rates (1.6% and 2.1% respectively). SIFT and RIFT have lower detection rates (87.0% and 87.3% respectively) with higher false-positive rates (3.8% and 5.1% respectively).

6.5.5 Summary of performance

We can now summarize and compare the performance of the four descriptors with the three assignment methods. Figure 6.21 shows how each descriptor performs for each assignment method. In Figure 6.21a we see the true positive detection performance where we observe the outperformance of density histogram (DH) and density-gradient histogram (DGH) against both SIFT and RIFT. In general the best



(a) True-positive performance

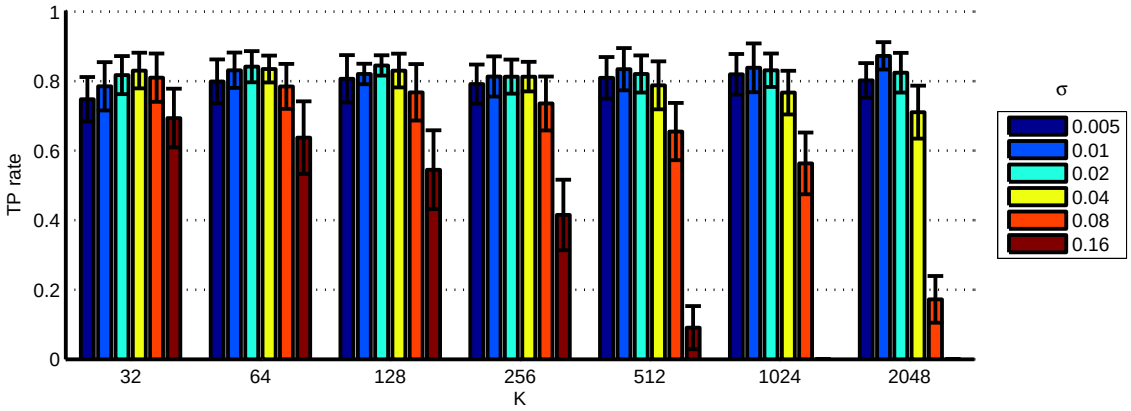


(b) False-positive performance

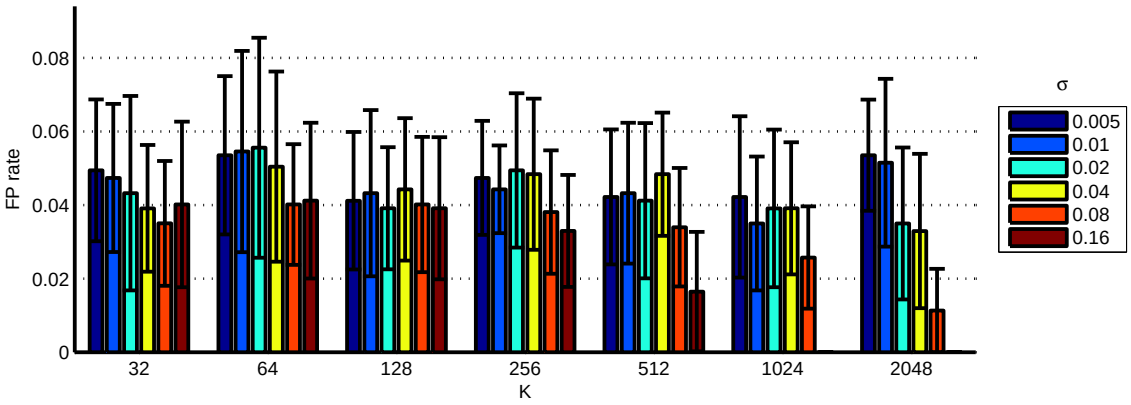
Figure 6.17: Handgun sub-volume results using uncertainty assignment, SVM classification for SIFT descriptor extending the range of smoothing parameter settings used

Descriptor	K	σ	TP rate (%)	FP rate (%)
SIFT	1024	0.02	87.0 ± 5.4	3.8 ± 2.4
RIFT	2048	0.01	87.3 ± 3.9	5.1 ± 2.3
DH	2048	0.02	97.2 ± 2.1	1.6± 1.4
DGH	512	0.04	97.2 ± 2.8	2.1 ± 1.3

Table 6.4: Best handgun detection rate for each descriptor using uncertainty assignment with SVM classifier

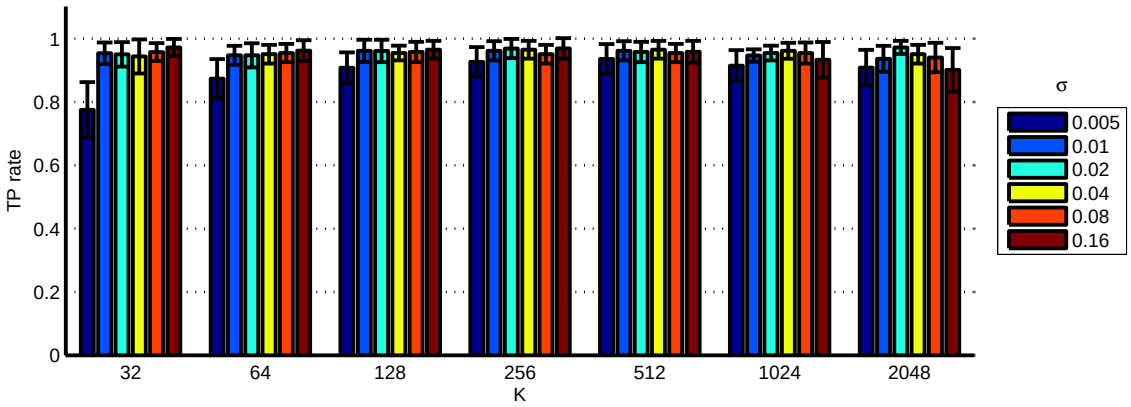


(a) True-positive performance

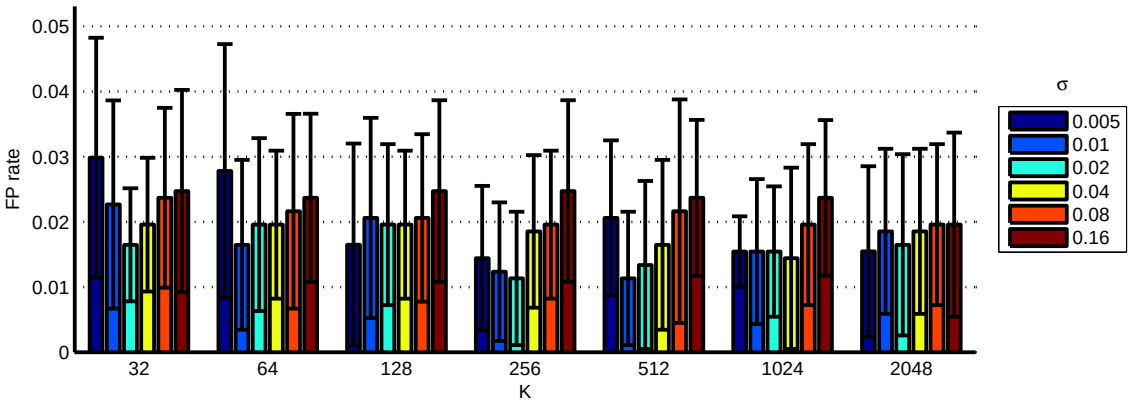


(b) False-positive performance

Figure 6.18: Handgun sub-volume results using uncertainty assignment, SVM classification for RIFT descriptor

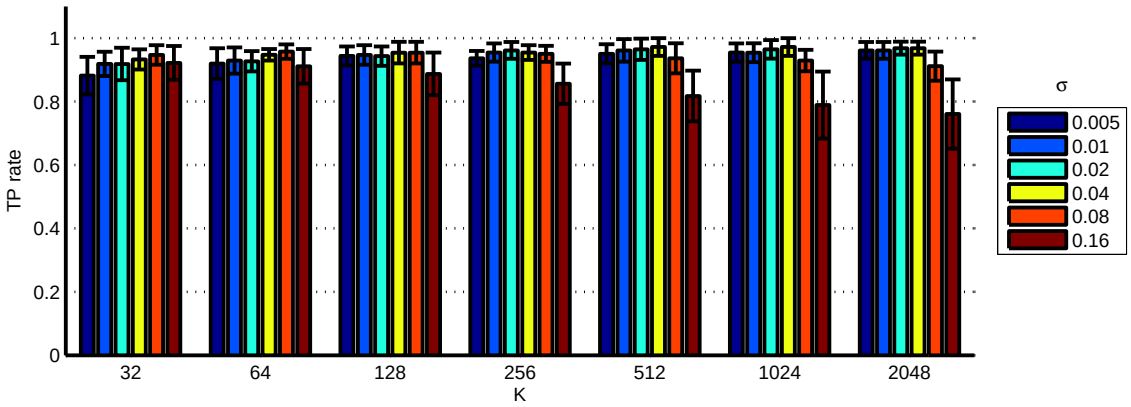


(a) True-positive performance

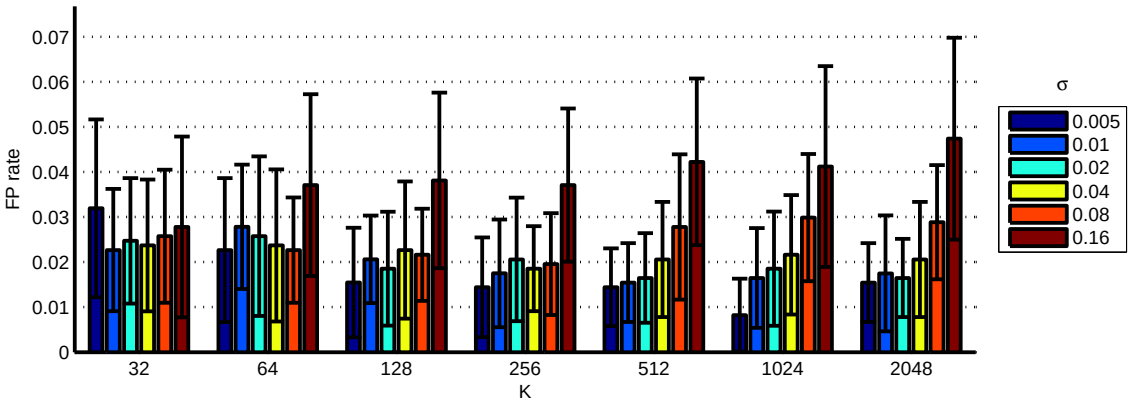


(b) False-positive performance

Figure 6.19: Handgun sub-volume results using uncertainty assignment, SVM classification for density-histogram descriptor

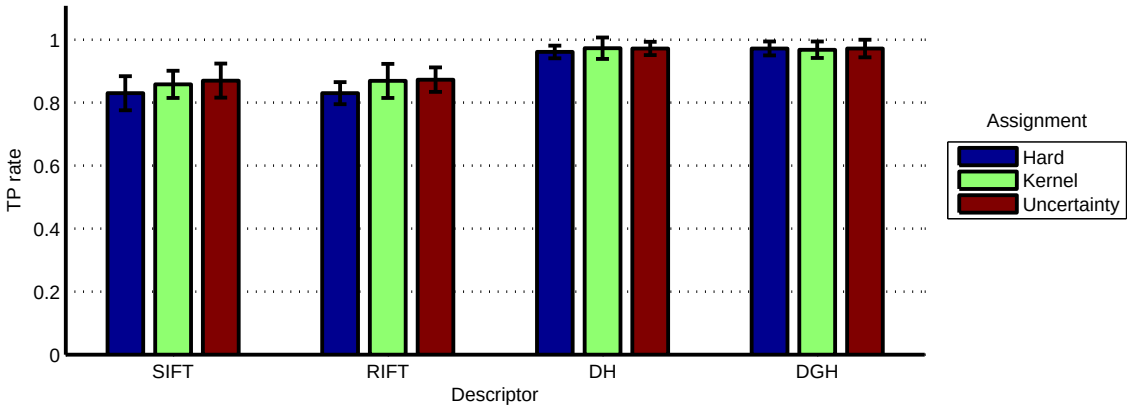


(a) True-positive performance

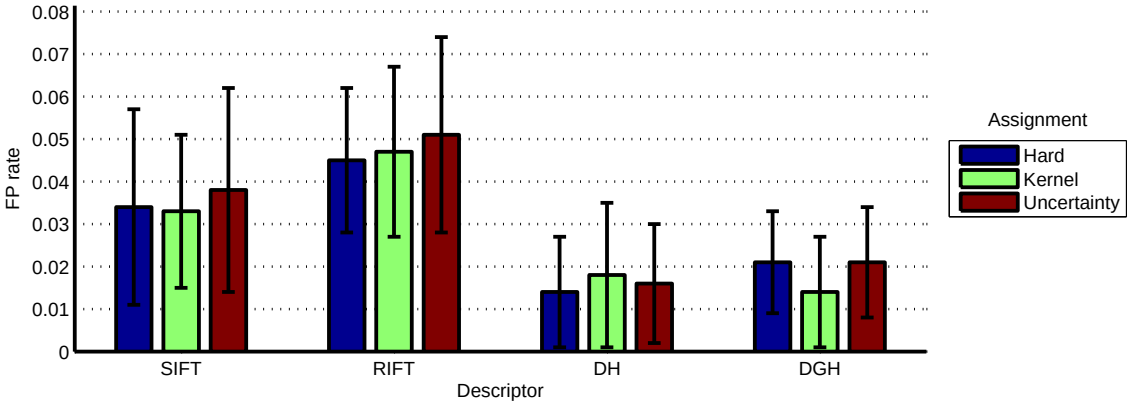


(b) False-positive performance

Figure 6.20: Handgun sub-volume results using uncertainty assignment, SVM classification for density-gradient histogram descriptor



(a) True-positive performance



(b) False-positive performance

Figure 6.21: Best handgun detection sub-volume results summary using SVM classification

detection is obtained using uncertainty assignment with hard assignment being the least effective in line with van Gemert et al. (2010), however the margin of error in these results does not allow this to be a clear cut conclusion and the performance of hard assignment for the DGH descriptor matches the best performance. Table 6.5 summarizes the best performing result for each descriptor where we can see that the density-histogram descriptor has the highest overall detection result (97.3%) with the lowest false-positive rate (1.8%). The density-gradient histogram is close behind but there is a marked difference to the SIFT and RIFT descriptors with detection rates of $\sim 87\%$ coupled with higher false-positive rates.

6.6 Results using bottle sub-volumes

Examination was also performed using bottles as the class of object to be recognized.

In the interests of brevity a detailed analysis of performance for each descriptor as

Descriptor	Assignment Method	K	σ	TP rate (%)	FP rate (%)	Precision	Recall
SIFT	Uncertainty	1024	0.02	87.0 ±5.4	3.8 ±2.4	0.870 ±0.069	0.870 ±0.054
RIFT	Uncertainty	2048	0.01	87.3 ±3.9	5.1 ±2.3	0.832 ±0.066	0.873 ±0.039
DH	Kernel	1024	0.04	97.3 ±3.4	1.8 ±1.7	0.942 ±0.053	0.972 ±0.034
DGH	Hard	2048	-	97.2 ±2.2	2.1 ±1.2	0.932 ±0.035	0.972 ±0.022
	Uncertainty	512	0.04	97.2 ±2.8	2.1 ±1.3	0.932 ±0.038	0.972 ±0.028

Table 6.5: Handgun sub-volume detection: best settings for each descriptor

the number of clusters and assignment method is varied are recorded in Appendix A. We will now present the settings and performance results that achieved the highest recognition rates.

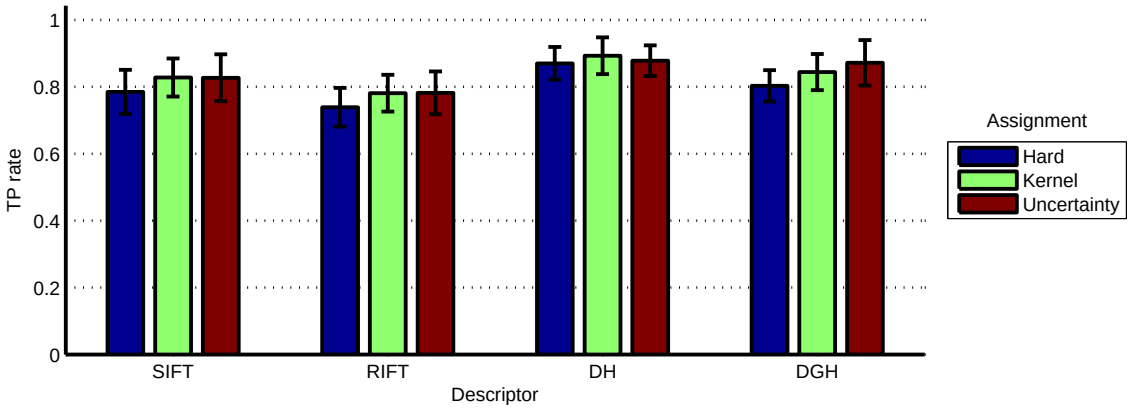
6.6.1 Summary of performance

We can compare relative performance for each assignment method and descriptor type by recording the settings that achieved peak detection rates. Figure 6.22a shows the peak true-positive performance for each descriptor as the assignment method is varied and Figure 6.22b shows the corresponding false-positive rates. The results are similar in nature to those obtained for handgun detection (Section 6.5.5) with hard assignment being the least effective methodology regardless of descriptor type. Uncertainty and kernel-assignment methods produce very similar results (within the measured error). Table 6.6 summarizes the parameter settings that achieve peak recognition results where we can see that uncertainty assignment gives best performance for three descriptors (SIFT, RIFT, DGH) with kernel assignment resulting in the highest overall detection rate of 89.3% for the DH descriptor. However, the measured error for all these results means a clear conclusion is not possible.

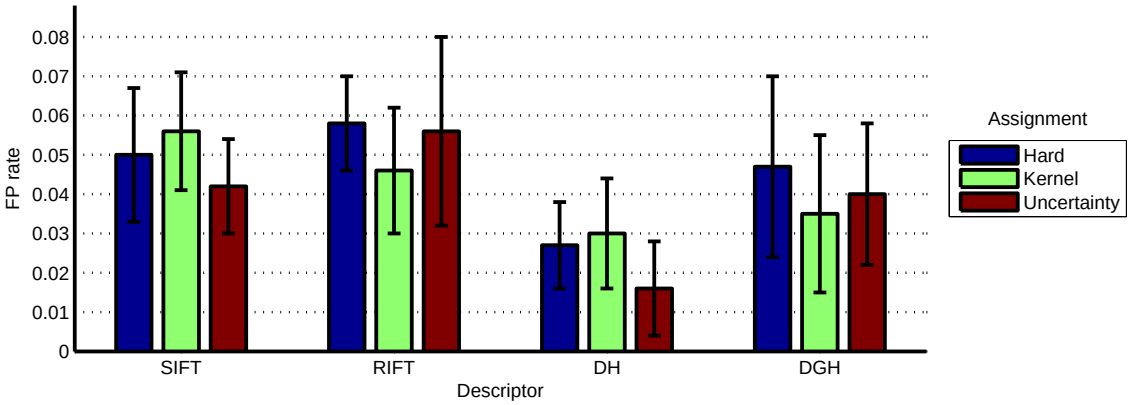
6.7 Interpretation of result data

6.7.1 Handgun recognition

In the classification of handgun sub-volumes the DH and DGH descriptors achieved high detection rates (in excess of 97.0%, see Section 6.5.5). It is worth examining



(a) True-positive performance



(b) False-positive performance

Figure 6.22: Best bottle detection sub-volume results summary using SVM classification

Descriptor	Assignment Method	K	σ	TP rate (%)	FP rate (%)	Precision	Recall
SIFT	Uncertainty	2048	0.02	82.7±7.0	4.2±1.2	0.900± 0.025	0.828± 0.070
RIFT	Uncertainty	2048	0.01	78.2±6.4	5.6±2.4	0.864± 0.052	0.783± 0.064
DH	Kernel	512	0.08	89.3±5.5	3.0±1.4	0.932± 0.029	0.893± 0.055
DGH	Uncertainty	512	0.04	87.2±6.8	4.0±1.8	0.908± 0.039	0.873± 0.068

Table 6.6: Bottle sub-volume detection: best settings for each descriptor

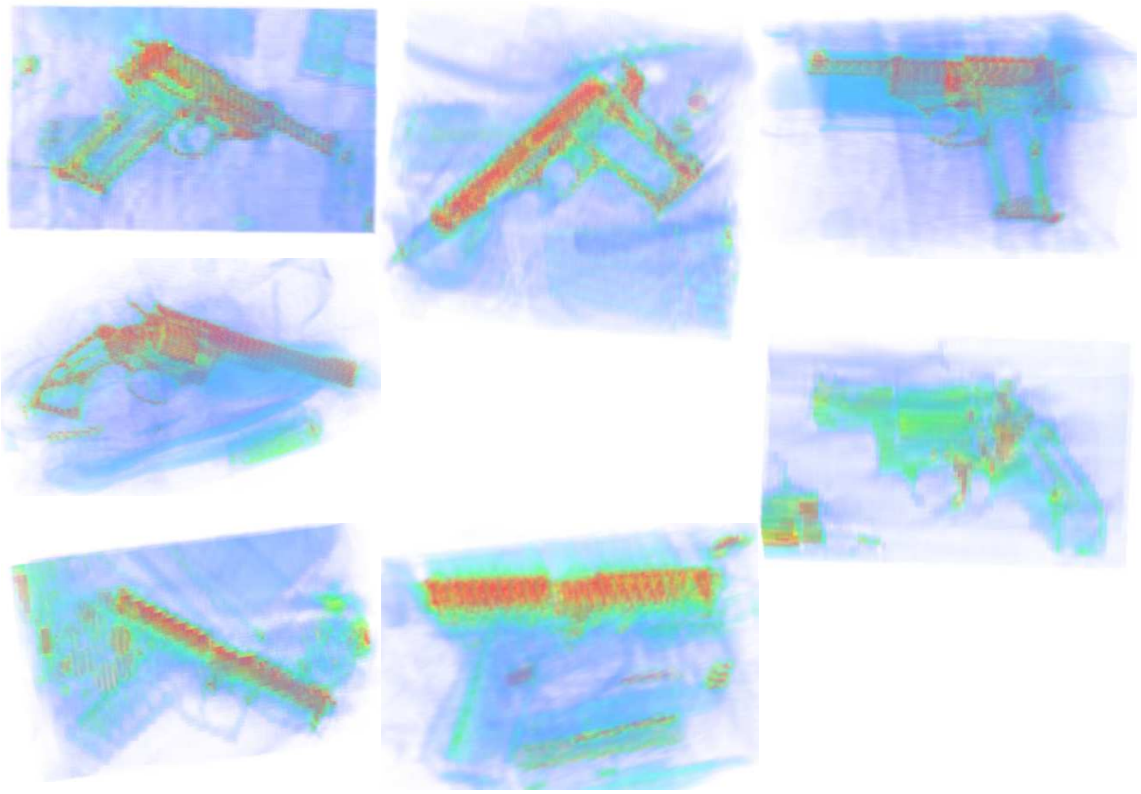


Figure 6.23: DH misclassification: missed handguns

the volumes that were classified in error to see if there is any obvious reason why some volumes were misclassified as clear or threat. We will examine these errors using the settings that achieved the highest detection rates (Table 6.5).

Figure 6.23 shows all seven handgun volumes that were misclassified as clear using the DH descriptor. There appears to be no obvious reason for the error (the same weapon consistently missed, a similar orientation for missed items, etc.). Figure 6.24 shows some of the clutter volumes that were misclassified as handguns. A number of these volumes contain batteries that appear to be triggering the misclassification. Other objects present include electrical transformers, inline roller skates and electronic equipment. All of these volumes contain some amount of metal.

DGH misclassification results are shown in Figure 6.25 and Figure 6.26. We can see the seven handguns that were not detected in Figure 6.25. Figure 6.26 shows some of the clutter volumes that were misclassified which show similar items as for the DH descriptor results: batteries, transformers, inline skates. This time there are volumes which do not contain metal: bottles, clothing and shoes.

SIFT misclassification results are shown in Figure 6.27 and Figure 6.28. The volumes that are misclassified as handguns (Figure 6.28) differ to the DH and DGH examples as there are now more volumes that contain little or no metal regions.

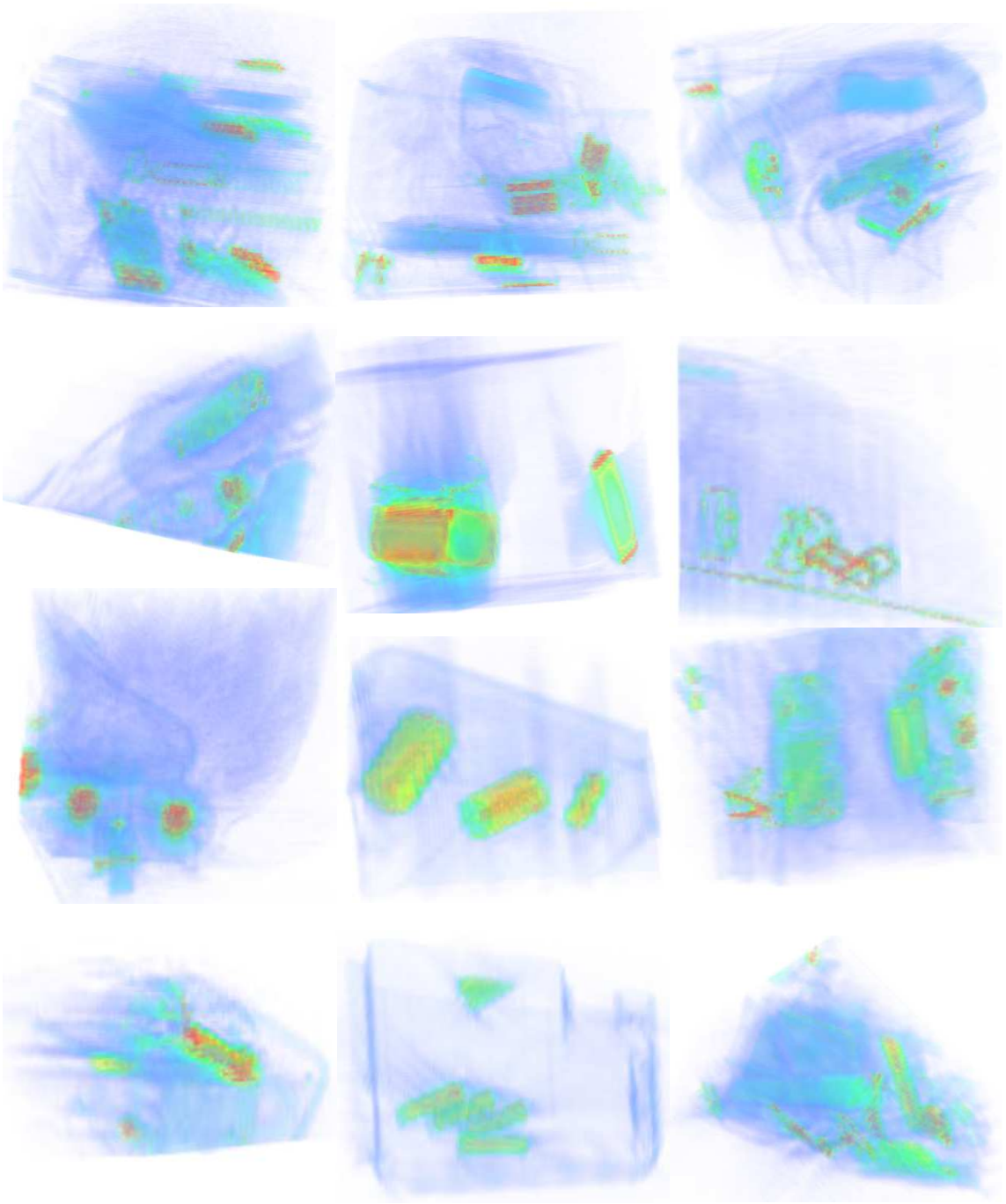


Figure 6.24: DH misclassification: clutter classed as handgun

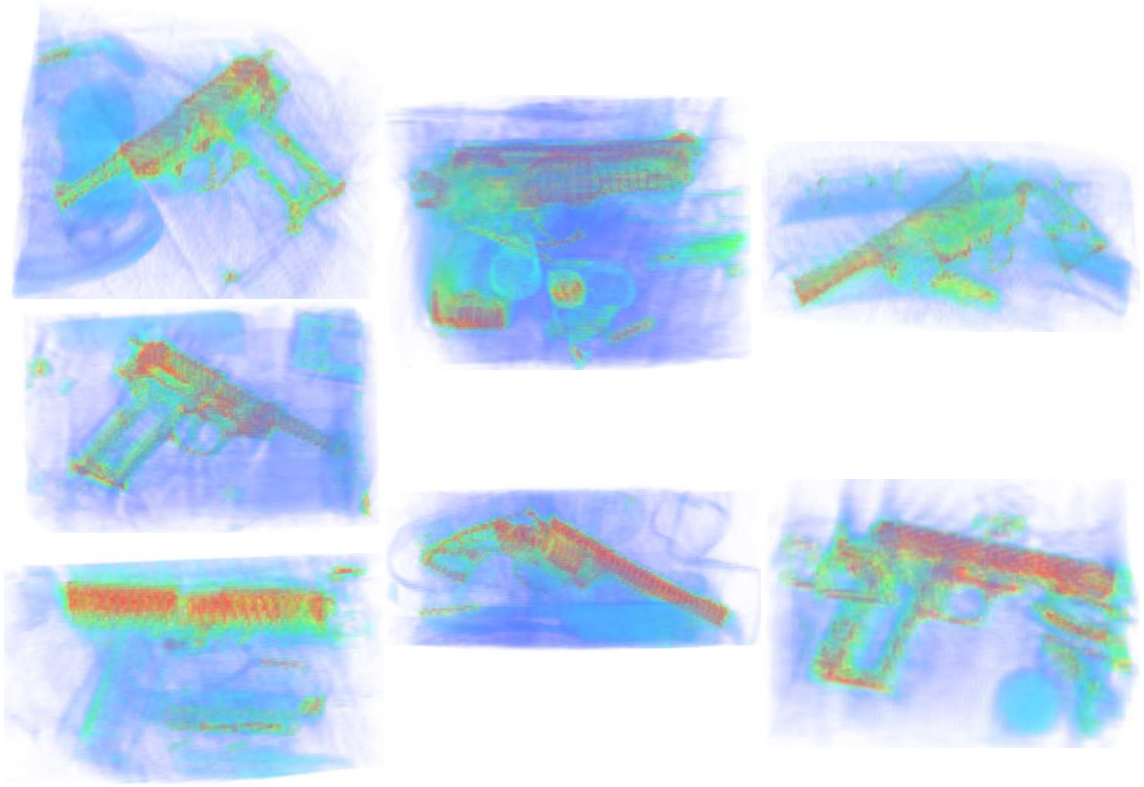


Figure 6.25: DGH misclassification: missed handguns

RIFT misclassification results are shown in Figure 6.29 and Figure 6.30. The volumes that are misclassified as handguns (Figure 6.30) are similar to the SIFT case in that there are now more volumes that contain few metallic regions.

Overall we can see that the choice of descriptor has a distinct impact on both the true-positive and false-positive recognition rates with the less complex DH and DGH descriptors outperforming the more complex and SIFT and RIFT descriptors. The choice of assignment methodology does influence the recognition results with 'soft' assignment (uncertainty or kernel) outperforming traditional hard assignment. It would appear that uncertainty assignment marginally outperforms kernel assignment when applied to the baggage CT data although the measured margins of error are not small enough to make this a significant claim. It is unclear from these results why some example images are not correctly identified; examination of the erroneous classifications has not yielded an obvious reason and this is left as an area for further work.

6.7.2 Bottle recognition

Recognition results obtained using bottle sub-volumes (Section 6.6, Appendix A) were lower than for handgun sub-volumes, indicating that this object class was more

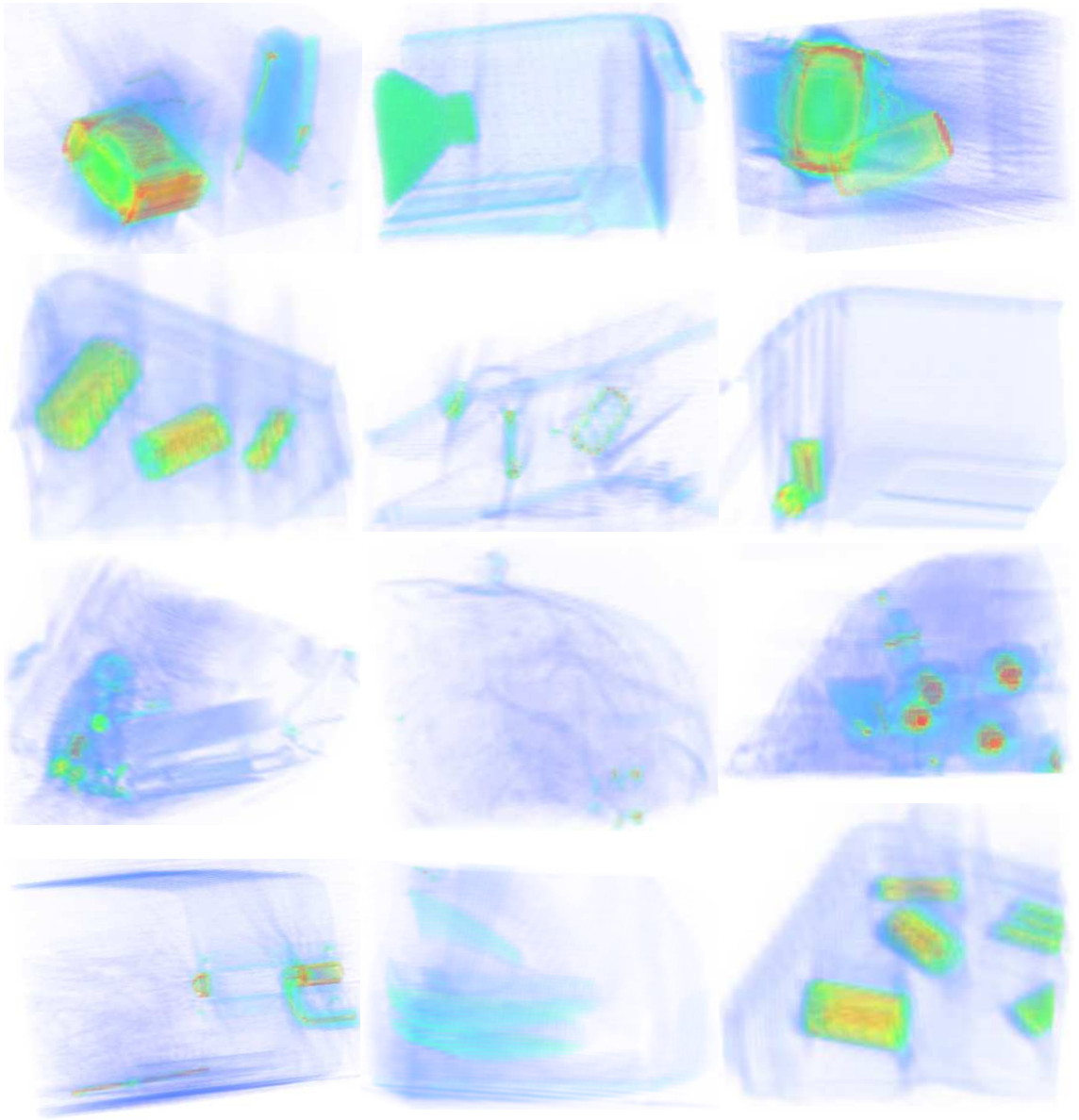


Figure 6.26: DGH misclassification: clutter classed as handgun

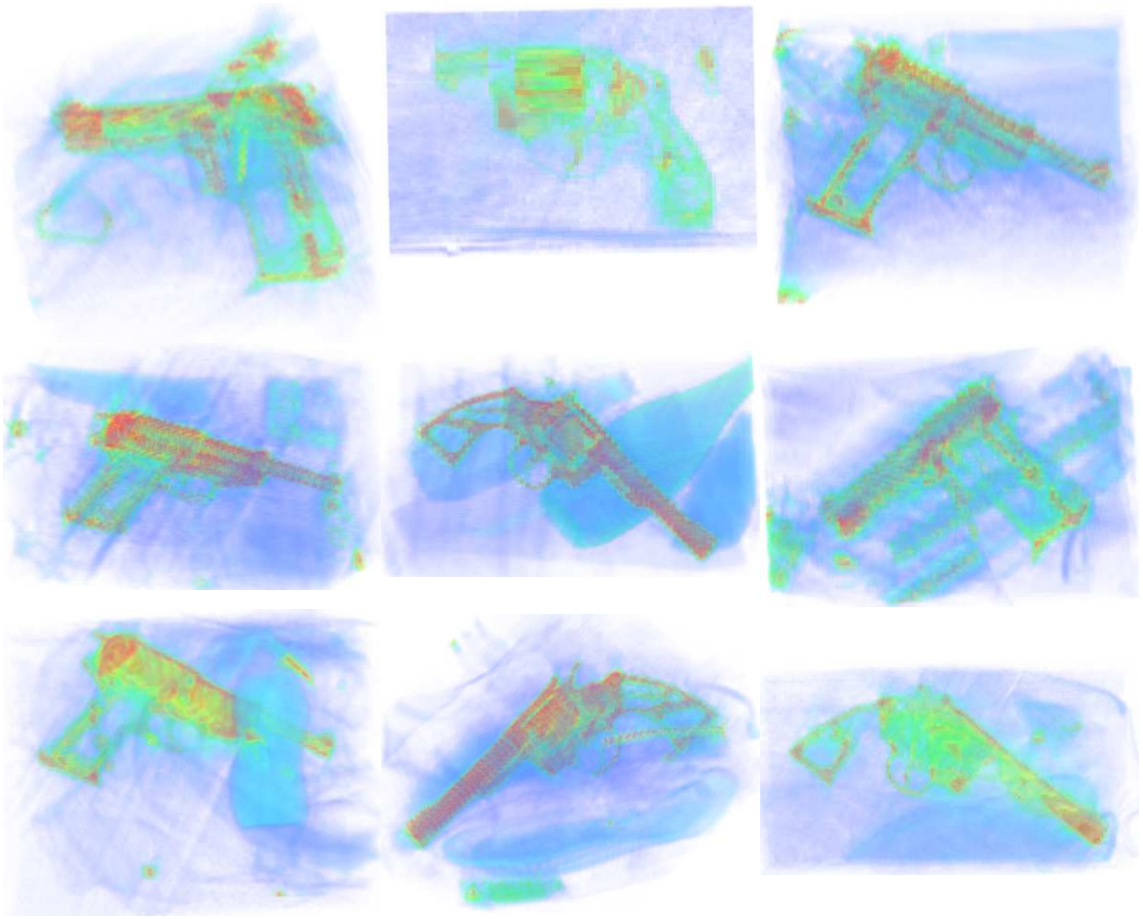


Figure 6.27: SIFT misclassification: missed handguns

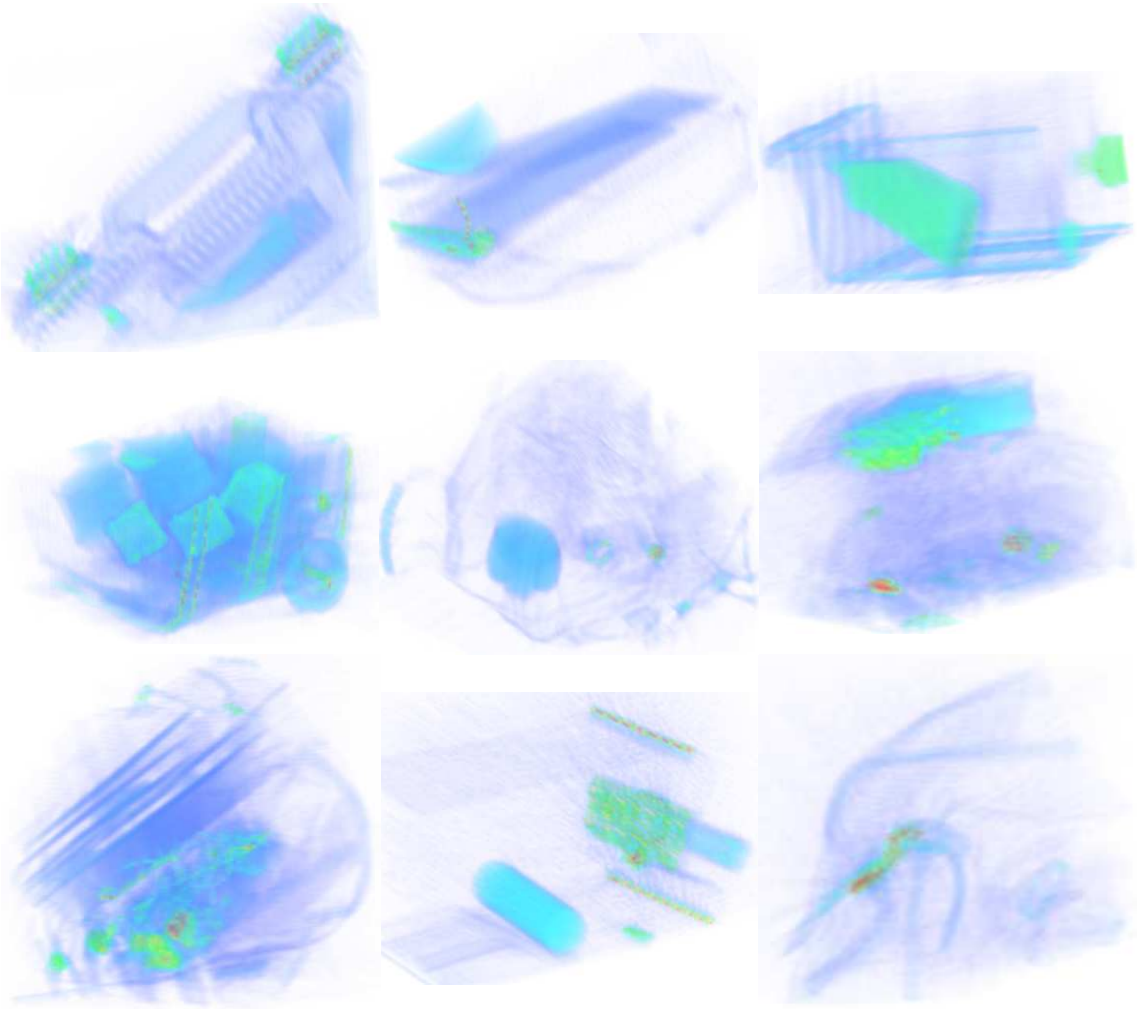


Figure 6.28: SIFT misclassification: clutter classed as handgun

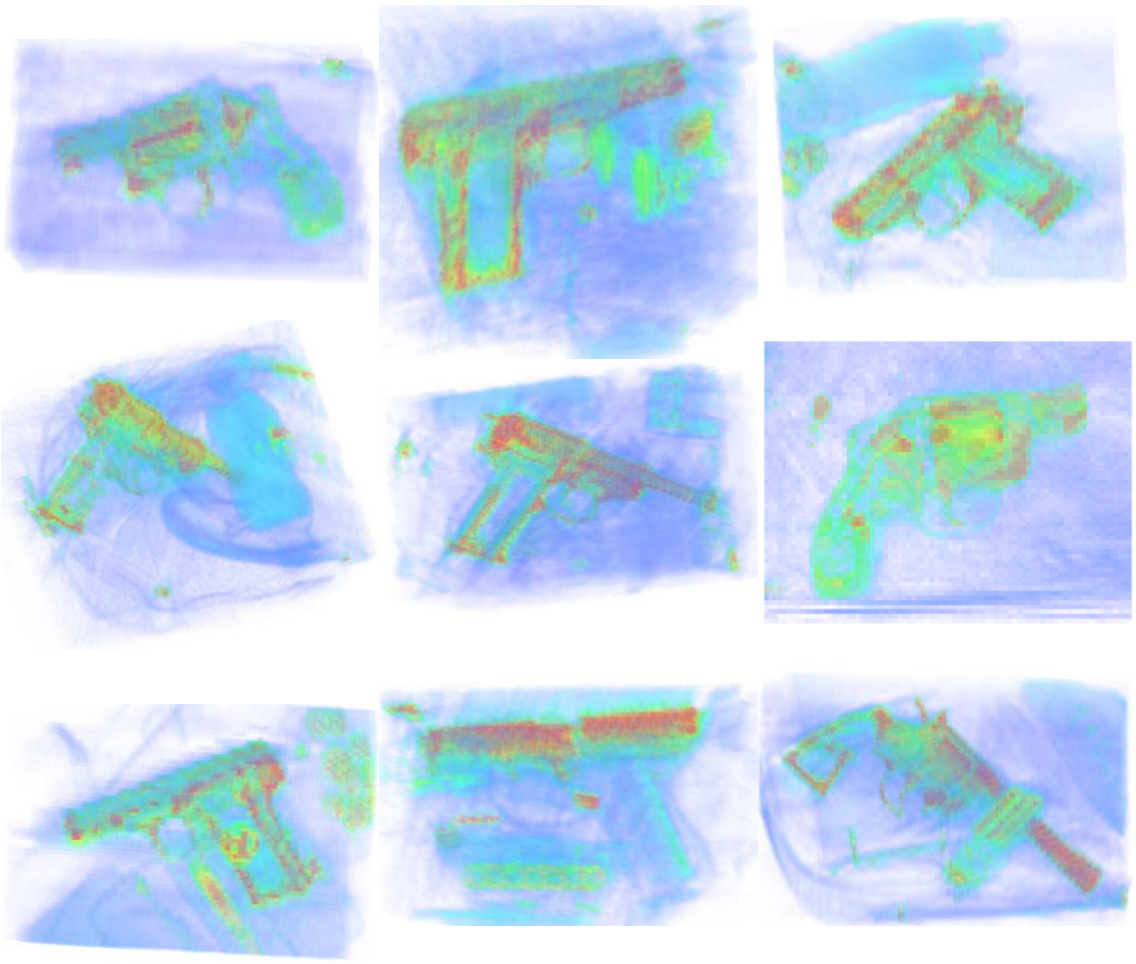


Figure 6.29: RIFT misclassification: missed handguns

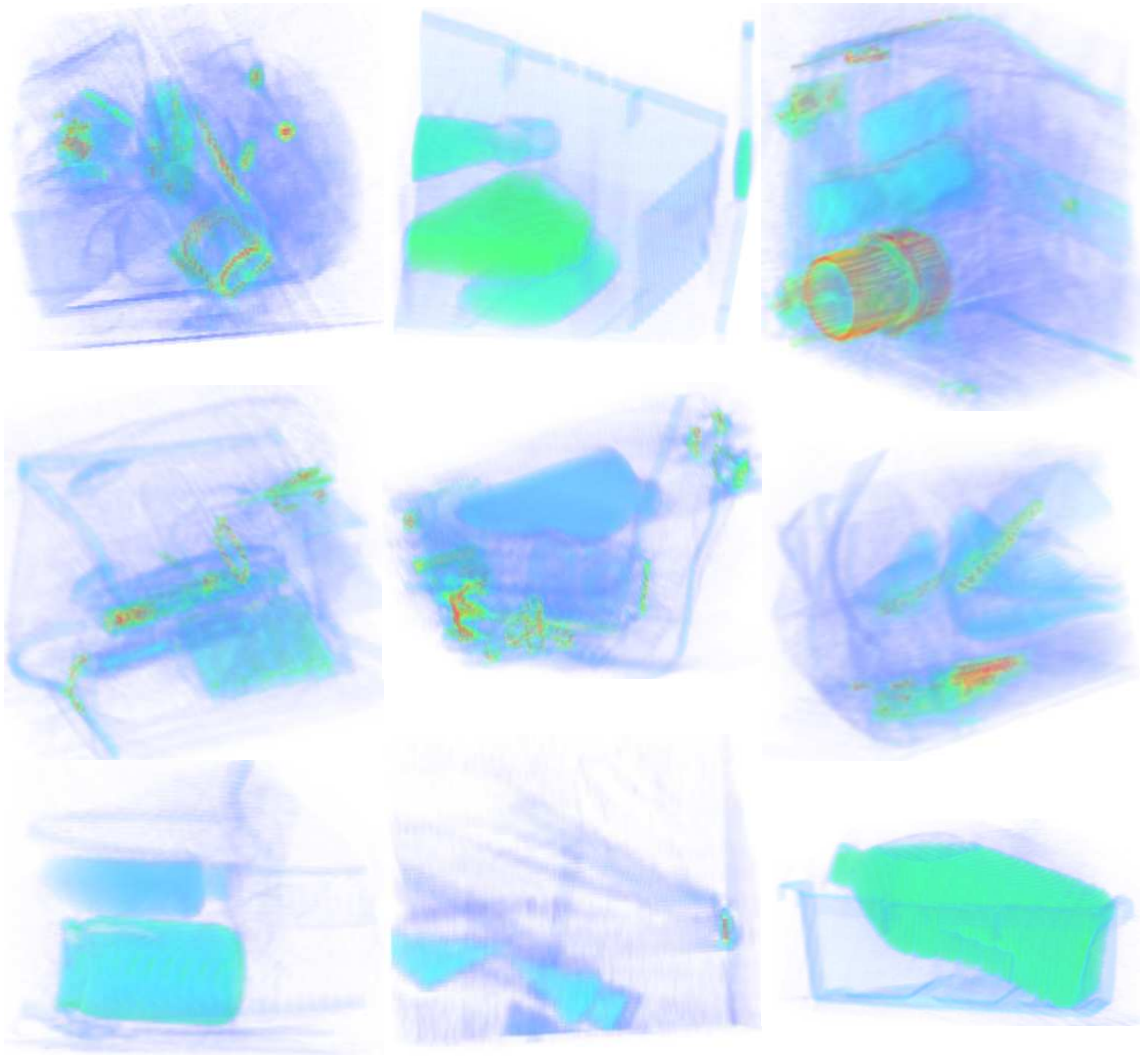


Figure 6.30: RIFT misclassification: clutter classed as handgun

difficult to recognize using the proposed methodology. We can again examine the misclassifications in an attempt to understand the reasons for the lesser performance.

Figure 6.31 shows examples of bottle sub-volumes that were misclassified as clutter for the DH descriptor in its optimum settings (Table 6.6). We can see bottles of various shapes and sizes containing varying degrees of liquid but no obvious reason for the misclassification. The misclassified bottles do not appear to be particularly challenging in nature. Figure 6.32 shows clutter that has been misclassified as a bottle. It can be seen that there are regions that are similar in density to the liquids used in the training set, which may be the cause of misclassification, but there is little evidence of 'bottle-shaped' items. This is one known flaw of the codebook approach: the geometric relationship between points of interest on an item are not considered within codebook entries (Lazebnik et al., 2006; Bosch et al., 2007).

Figure 6.33 shows examples of bottle sub-volumes that were misclassified as clutter for the DGH descriptor in its optimum settings (Table 6.6). We can again see a range of bottle items including some that may be considered challenging (virtually empty deodorant bottles). Figure 6.34 shows clutter that has been misclassified as a bottle where we can see some items that could contain gradients similar to those from genuine bottle objects. In particular we can see some perspex rods that have been misclassified (although not all instances of this item were misclassified).

Figure 6.35 shows examples of bottle sub-volumes that were misclassified as clutter for the SIFT descriptor in its optimum settings (Table 6.6). Again we can see bottles with a variety of poses, shapes and sizes containing varying degrees of liquid. There appears to be no common feature of these volumes that would indicate a reason for their misclassification. Figure 6.36 shows clutter that has been misclassified as a bottle when the SIFT descriptor is used. It is noticeable that more metallic objects appear in this dataset some of whose features, when normalized during the SIFT descriptor generation, may be similar to bottles in shape: some electrical transformers are present whose circular cross-section is similar to that of a full bottle. We can also see some batteries (again with a circular cross-section) and the perspex rods.

Figure 6.37 shows examples of bottle sub-volumes that were misclassified as clutter for the RIFT descriptor in its optimum settings (Table 6.6). We again see a selection of bottle items from the dataset that exhibit varying pose, size, shape and liquid content. These misclassifications show less metal than the SIFT misclassifications but there is no obvious mode of failure in the misclassification other than the possibility that the RIFT descriptor is not resulting in a codebook that distinctly characterizes bottles. Figure 6.38 shows clutter that has been misclassified

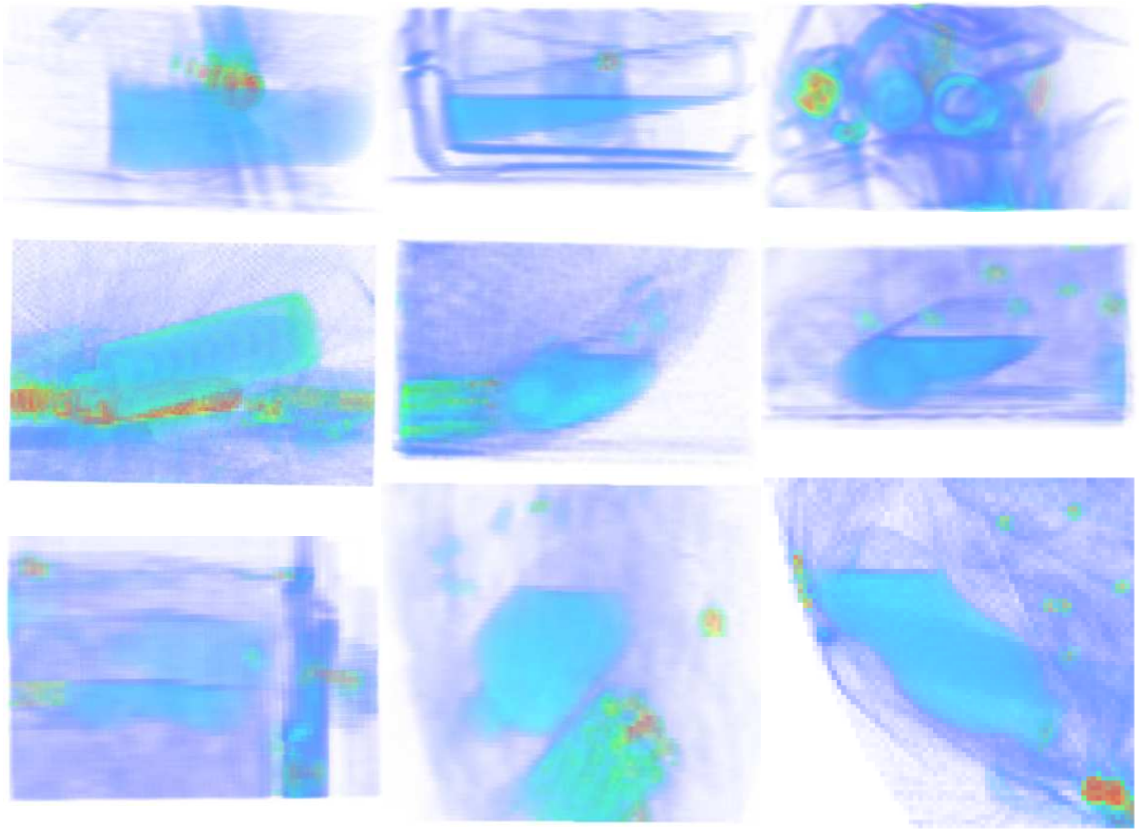


Figure 6.31: DH misclassification: missed bottles

as a bottle which shows a range of volumes, some showing metallic objects, others containing little other than a plastic tray used to hold items as they transit the CT scanner. These images show little that resembles a bottle in nature.

6.8 Results using handgun whole volumes

Experiments were also performed using whole-baggage volumes for the dataset. In the interests of brevity, we present classification results for the handgun class as an exemplar. For this case a whole bag was either marked as "clear" if a handgun was not contained or "threat" if a handgun was present.

We now present a concise summary of the results. Detailed analysis is given in Appendix B.

6.8.1 Summary of performance

We summarize and compare the performance of the four descriptors with the three assignment methods in the analysis of whole baggage items. Figure 6.39 shows how each descriptor performs for each assignment method when we take the configura-

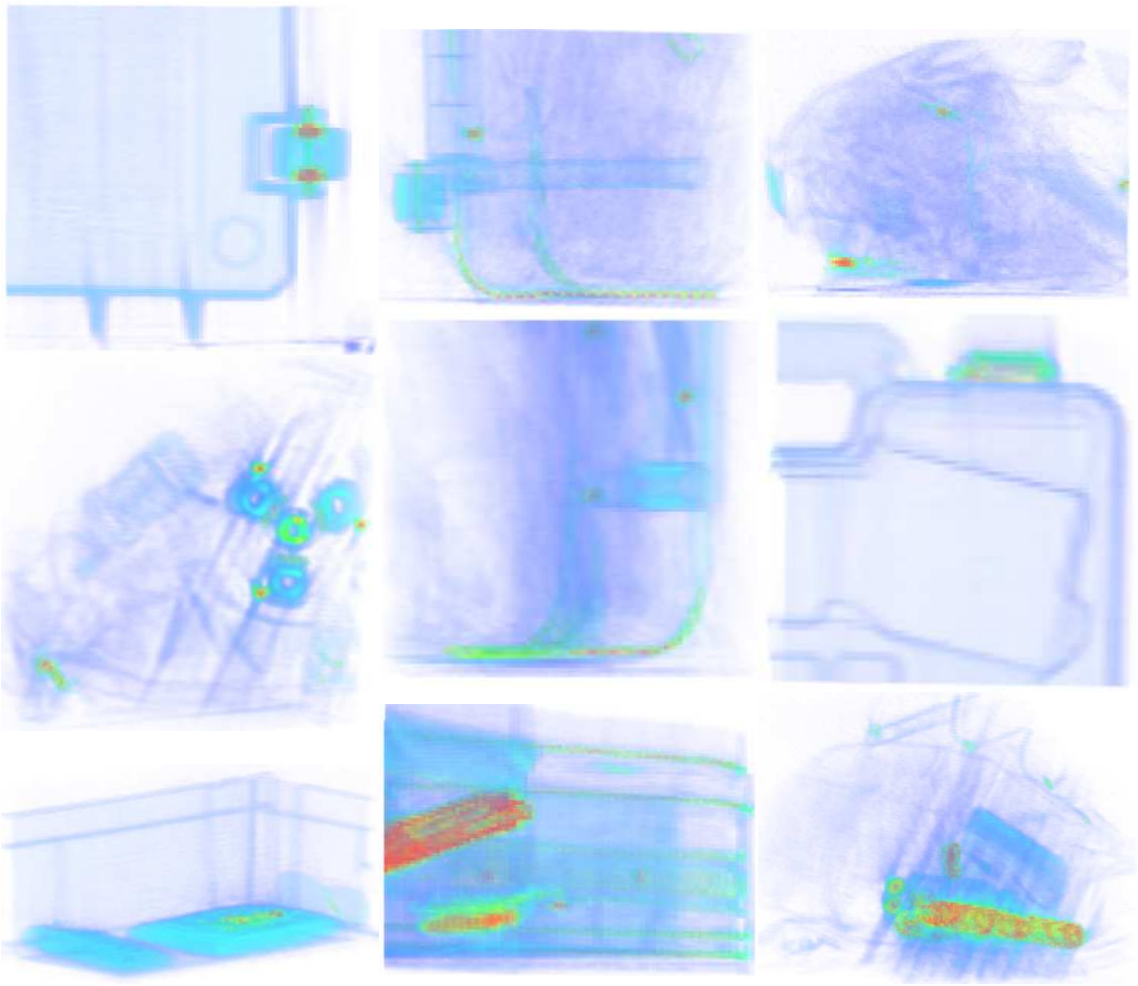


Figure 6.32: DH misclassification: clutter classed as bottle

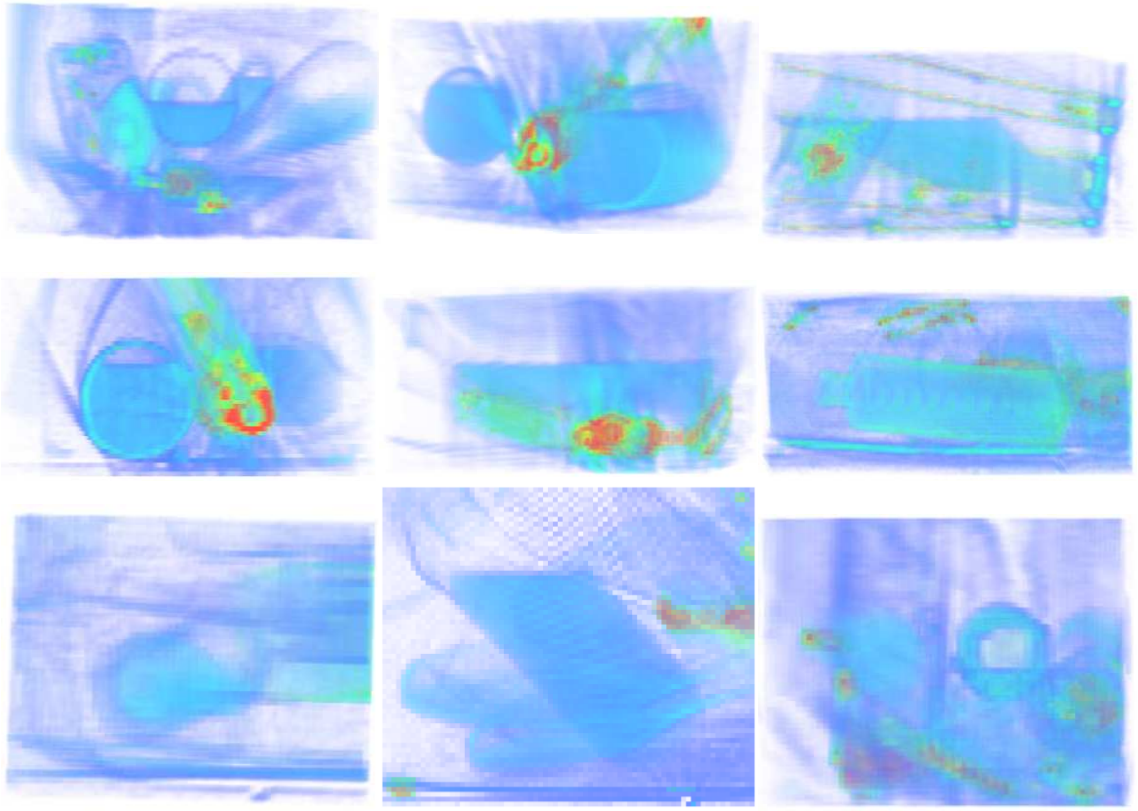


Figure 6.33: DGH misclassification: missed bottles

tion that yields the highest true-positive result. Figure 6.39a shows the detection performance where we can see the general outperformance of density histogram over density-gradient histogram, SIFT and RIFT. The best detection is obtained using kernel assignment with the density-histogram descriptor: 94.8% true-positive rate. This setting also yields a relatively low false-positive result: 15.7%. SIFT and RIFT have significantly higher false-positive rates when compared to DH and DGH. Table 6.7 summarizes the best performing result for each descriptor where we can see that the density-histogram descriptor has the highest overall detection result (94.8%) with almost the lowest false-positive rate (15.7%).

The main observation from this work was that all four descriptors performed similarly in term of true-positive rate ($> 90.0\%$) with the DH descriptor yielding the best rate (94.8%). The main difference is now in the false-positive rate. The DH and DGH descriptors yield similar false-positive rates ($\approx 15\%$) which is in sharp contrast to that obtained for the SIFT descriptor ($\approx 45\%$) and RIFT descriptor ($\approx 56\%$). A number of observations can be made following these results. A comparison can be made to the handgun sub-volume results (summarized in Table 6.5) where we can see the DH and DGH descriptors clearly outperforming the SIFT and RIFT descriptors regarding true-positive rate - a result that is not as distinct in the whole

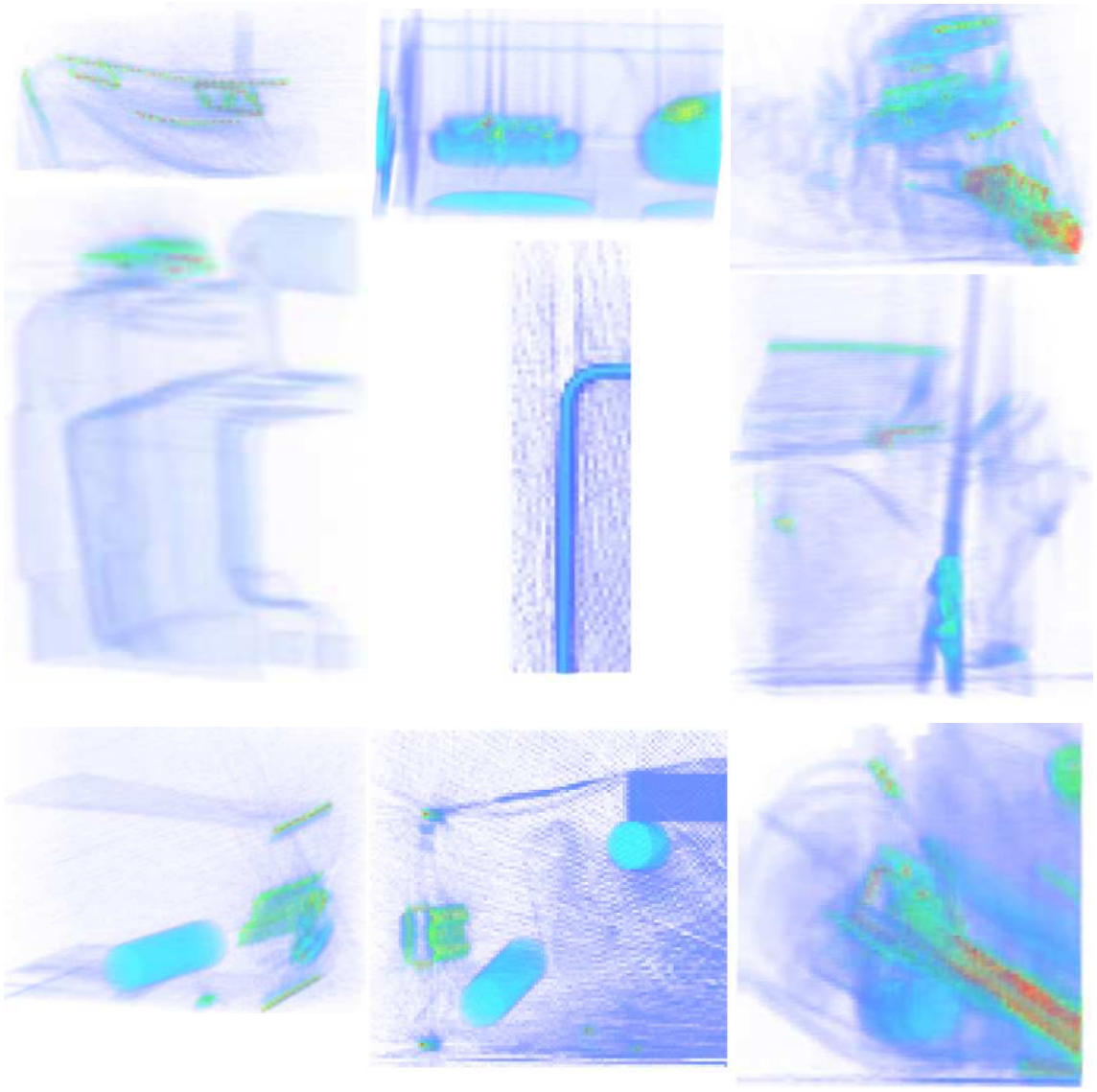


Figure 6.34: DGH misclassification: clutter classed as bottle

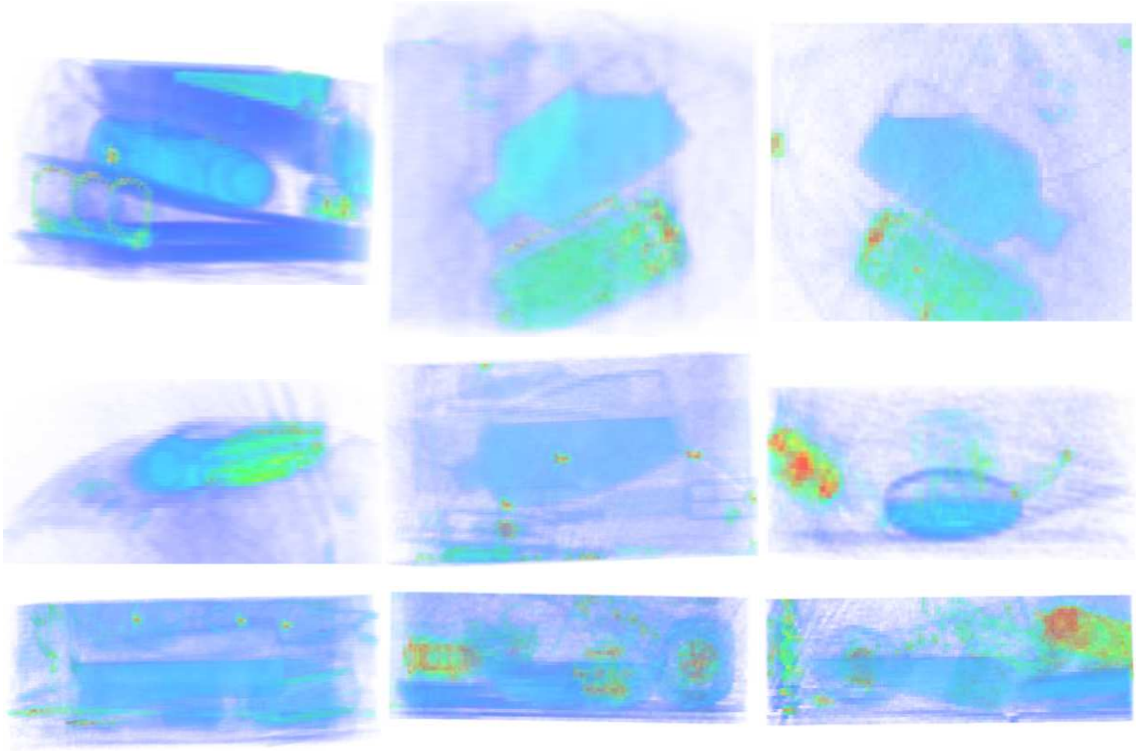


Figure 6.35: SIFT misclassification: missed bottles

volume dataset. We can also see that the difference in false-positive performance is a common aspect between the sub-volume and whole volume measurements. The high false-positive rates for SIFT and RIFT indicate that the true-positive results for these cases are being raised, not by a decision that a handgun is present, but by the increased likelihood that clutter within a whole bag will be classified as a handgun. The higher false-positive rates for all descriptors when compared to the sub-volume case is a natural extension of the increased baggage complexity - the handgun sub-volumes contain clutter but not to the same degree as that presented by a whole-baggage volume. The false-positive rates for whole baggage items mirror those obtained for sub-volumes with DH and DGH descriptors clearly outperforming SIFT and RIFT. It should be noted that, due to the size of the dataset, the measurement error is relatively large in each case - larger datasets are required to improve this aspect of the work.

This work is presented to show the application of the bag-of-features technique to whole-baggage volumes. Further analysis of whole baggage items containing bottles is an area of work that is left for the future.

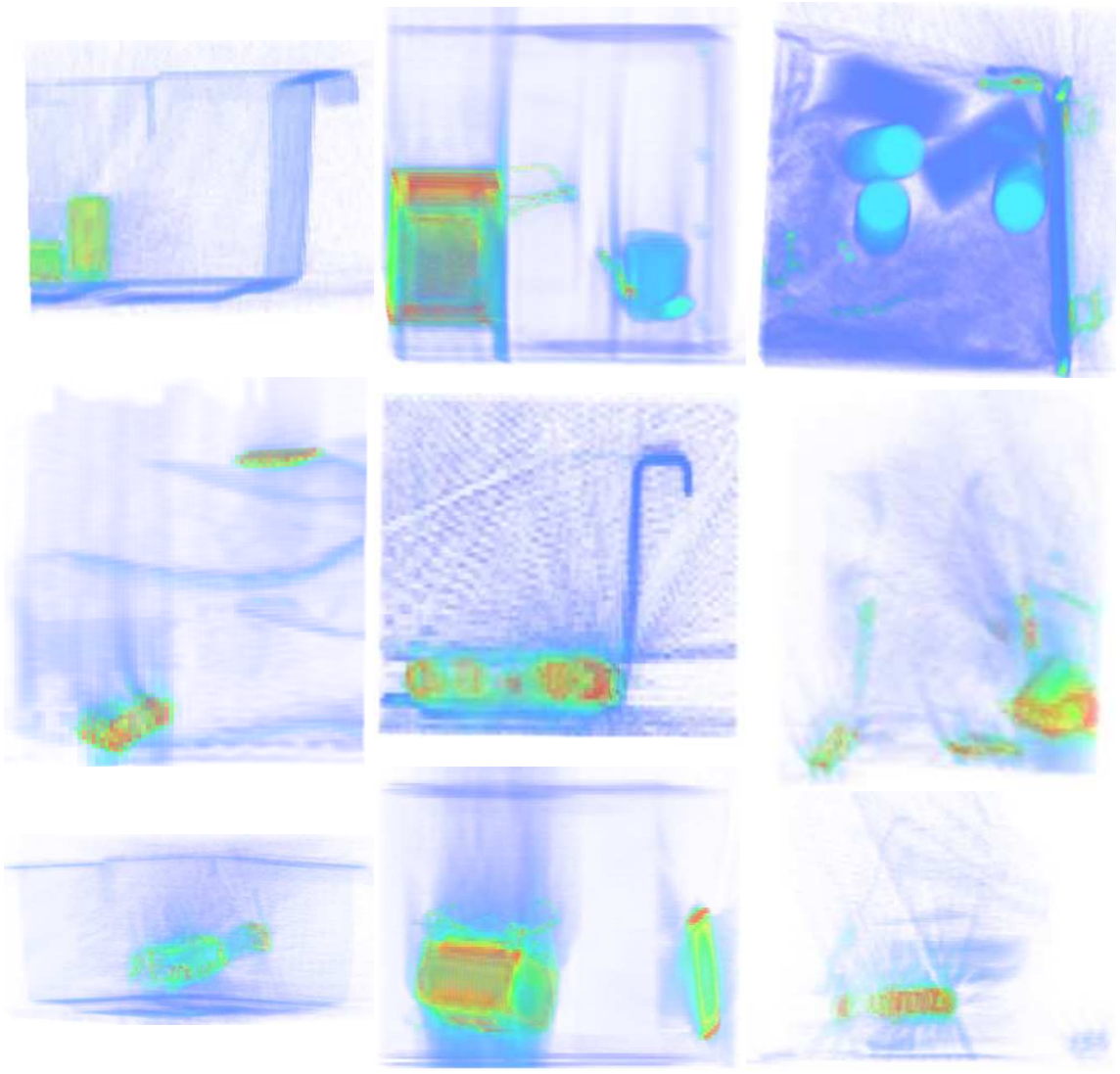


Figure 6.36: SIFT misclassification: clutter classed as bottle

Descriptor	Assignment Method	K	σ	TP rate (%)	FP rate (%)
SIFT	Uncertainty	2048	0.02	92.3 ± 5.4	44.8 ± 8.5
RIFT	Kernel	64	0.16	90.7 ± 6.5	56.4 ± 9.7
DH	Kernel	1024	0.16	94.8 ± 3.7	15.7 ± 12.6
DGH	Kernel	2048	0.08	91.2 ± 6.6	15.3 ± 8.8

Table 6.7: Whole-volume handgun best detection results and parametric settings

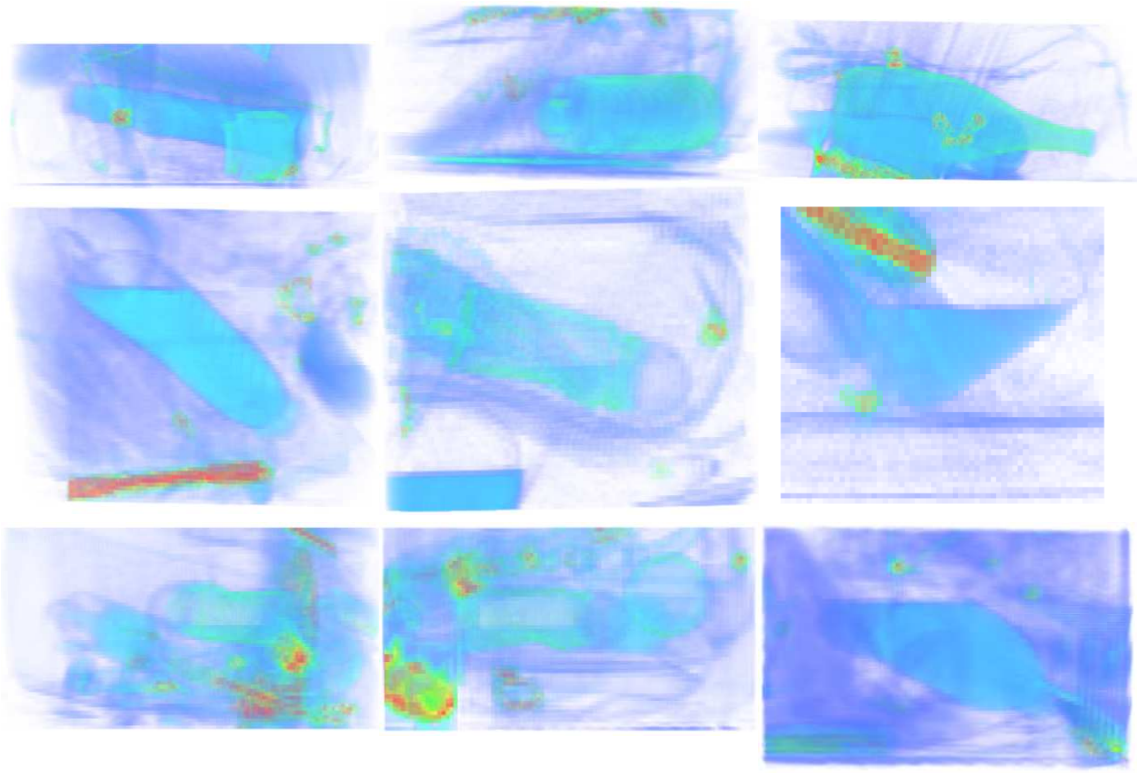


Figure 6.37: RIFT misclassification: missed bottles

6.9 Conclusions

This work has explored the use of the bag-of-features approach in the classification of two classes of threat item in CT-baggage imagery. In the case of handguns we have demonstrated high detection rates for both sub-volume and whole volume datasets. The density-histogram descriptor achieved the highest detection rates (97.3% for handgun sub-volumes; 89.3% for bottle sub-volumes) closely followed by the density-gradient histogram (97.2% for handgun sub-volumes; 87.2% for bottle sub-volumes). These descriptors also produced the lowest false-positive rates (DH: 1.8%, DGH: 2.1% for handgun sub-volumes; DH: 3.0%, DGH: 4.0% for bottle sub-volumes). The performance of the SIFT and RIFT descriptors was poor in comparison with lower detection rates and higher false-positive rates. Detection of a handgun within a whole volume image (rather than a sub-volume) again showed the density-histogram descriptor yielding the highest recognition rate (94.8%). This result can be compared to that obtained in Chapter 4 and Chapter 5 where we examined specific instance recognition within whole-baggage volumes. We see a similar relative performance between the descriptors (RIFT and SIFT are less effective than DH and DGH) but it is interesting to note that the true-positive rates obtained for class recognition (for a given false-positive rate) are similar given the more complex nature of object-class

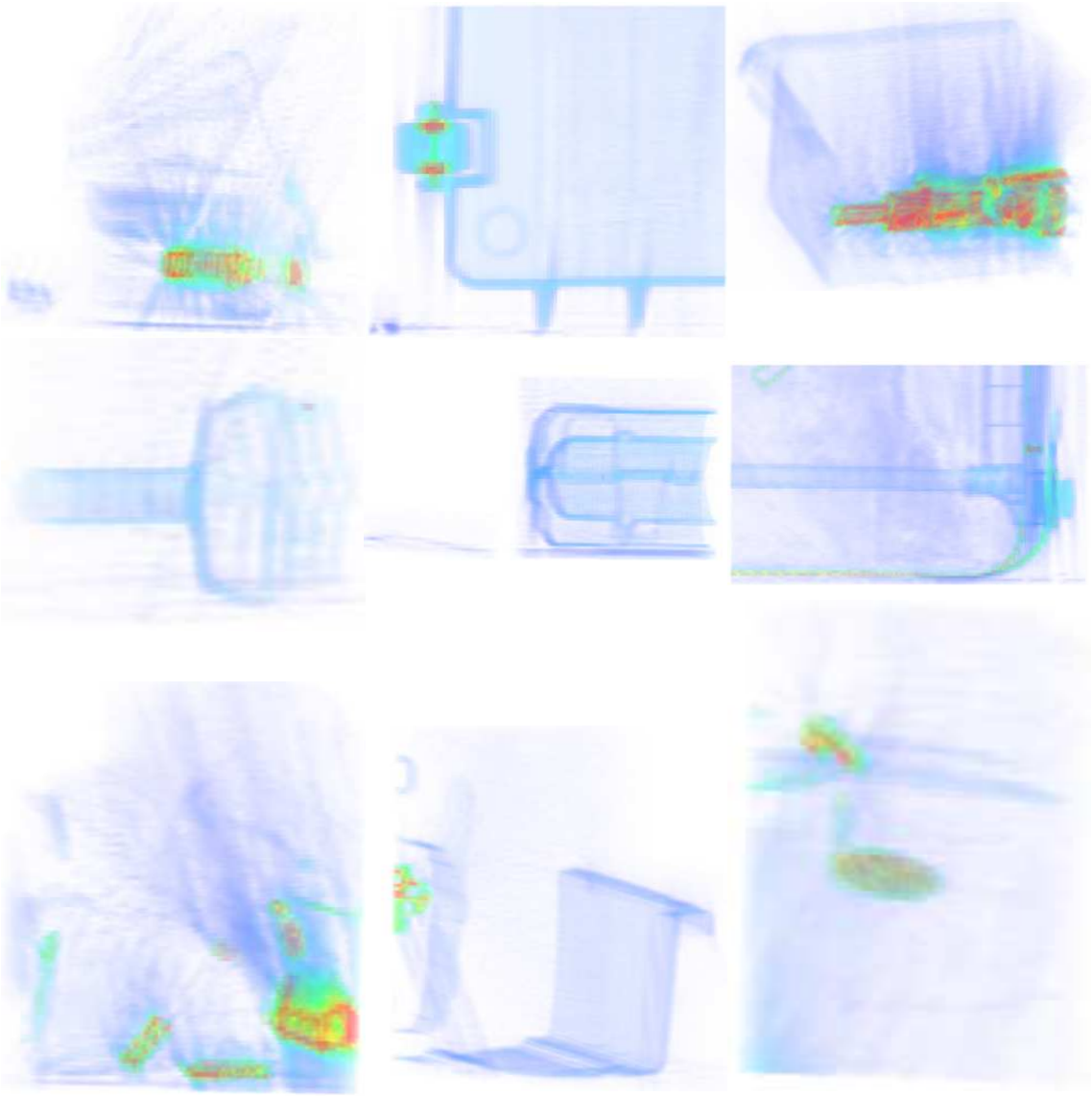
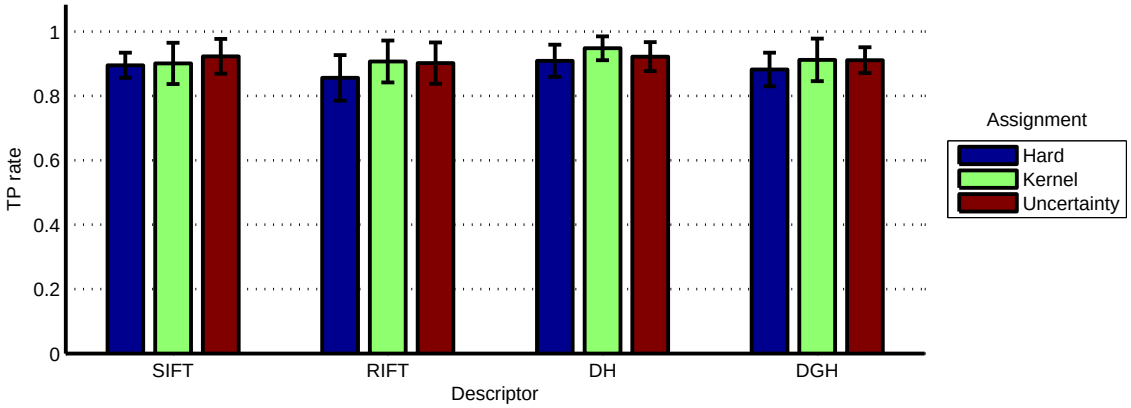
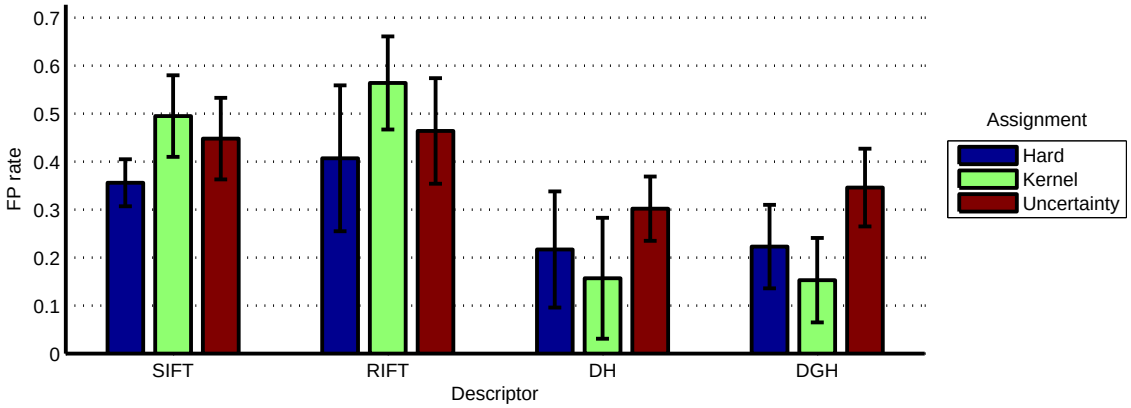


Figure 6.38: RIFT misclassification: clutter classed as bottle



(a) True-positive performance



(b) False-positive performance

Figure 6.39: Best detection whole-volume handgun results summary using SVM classification

recognition.

Recognition of bottle objects lags behind that of handguns which can attributed to the relative featureless nature of bottles and their contents. Liquids present as a large volumetric region of near constant density that will result in few points of interest (see Section 4.2). It is somewhat surprising that the recognition results are relatively high (>85.0%). We speculate that points of interest may be created from the imaging artefacts and thus allow description of the liquid regions.

The assignment methodology (hard, kernel, uncertainty) did influence performance, in part replicating the findings of van Gemert et al. (2010). For sub-volume handguns and bottles the best performance was, in general, achieved using uncertainty assignment, closely followed by kernel assignment. For the whole-volume handguns dataset the opposite was true: kernel assignment narrowly outperformed uncertainty assignment, though the error margins are such that this is not a definitive conclusion. Hard assignment lagged behind in all cases as demonstrated by van Gemert et al. (2010).

The choice of the number of visual words used (K) showed some influence on the results with best performance being achieved for $K = \{512, 1024, 2048\}$, a result also replicated by van Gemert et al. (2010), that indicates too few visual words result in a codebook lacking in sufficient salient entries to accurately describe the baggage volume.

The high margins of error, as given by the standard deviation results, are primarily due to the small dataset size. For example, each test dataset for handgun sub-volumes has typically 28 handguns and 97 clutter volumes leading to a resolution of 3.6% for true-positive results and 1.0% for false-positive results in each cross-validation test. An increase in the dataset size for all classes of object would be a desirable feature of future work. The standard deviation results are higher, relative to the quoted means, for the false-positive results. It is unclear why this would be the case but would need to be investigated on future analysis.

Recognition of handguns within whole-baggage volumes (Section 6.8, Appendix B) showed high true-positive rates (DH: 94.8%) but correspondingly higher false-positive rates (DH: 15.7%). This demonstrates a relationship between sub-volume results and whole-volume results. In our work volumetric data within the whole-volumes dataset are typically an order of magnitude greater in volume than those in the sub-volume dataset. Comparison of the false-positive results for whole volumes (Table 6.7) and sub-volumes (Table 6.5) demonstrate a corresponding order of magnitude difference. For example, density histogram false-positive results are 1.8% for sub-volumes and 15.7% for whole volumes; SIFT false-positive rates are 3.8% for

sub-volumes and 44.8% for whole volumes.

Analysis of misclassifications did not reveal any consistent errors: no particular gun, bottle or item of clutter was regularly misclassified.

Future work could include threat localization within a whole-baggage volume through scanning as a series of sub-volumes in a similar fashion to a traditional 2D approach (Viola and Jones, 2001; Dalal and Triggs, 2005; Mutch and Lowe, 2008). A comparison between a sub-volume window approach and a whole volume 'all in one' method (Section 6.8, Appendix B) for object recognition and an exploration of methods to improve the overall recognition performance in each case is an area of interest. It would also be useful to extend the results with the generation of ROC plots, as in Chapter 5, to examine the trade-offs that can be made between true-positive and false-positive rates.

This work has created a set of volumes that can act as a test bed for the comparison to alternative recognition approaches. The cross-validation method employed typically trains the classifier from a dataset comprising ≈ 1000 sub-volumes. One alternative approach is the extension of visual cortex-modelling methods into 3D that have been shown to produce excellent results with few training examples (Section 2.3). We shall examine visual cortex modelling and performance in the next chapter.

Chapter 7

A visual-cortex approach to object detection

By way of contrast to the interest-point approaches used before we now examine a method based on modelling functional aspects of the human visual cortex. A novel extension of an existing 2D model into 3D is presented for the recognition of handguns and liquid containers.

7.1 Introduction

The visual cortex is the region of the brain responsible for processing visual information arriving from the retina via the thalamus. The visual cortex has been intensively studied in various mammalian brains (cat, macaque monkey, spider monkey) as a means to derive functional models (Hubel and Wiesel, 1959, 1962, 1968). A subregion of the visual cortex called the primary visual cortex (V1) is the most studied area. It was discovered that V1 is hierarchical in structure with Simple (S) and Complex (C) neurons forming the basis of the hierarchy (Riesenhuber and Poggio, 1999). The work of Hubel and Wiesel (1959, 1962, 1968) examined V1 functionality by probing the visual cortex of unconscious cats and monkeys with electrodes and noting the electrical response in the cortex as differing visual stimuli were presented to the eye. This revealed that some of the simple neurons in the V1 region respond to oriented bars and edges, which led to the use of Gabor filters of varying orientation being used in software models (Serre et al., 2005b; Mutch and Lowe, 2008; Jhuang et al., 2007). As we move up through the hierarchy in the visual cortex, the cortex response is increasingly invariant to object transformations (scale, position) whilst also becoming focussed on more specific features (relevant to objects of interest). These processes have been modelled through the use of "max-pooling" operations

(Riesenhuber and Poggio, 1999) and comparison with learnt salient patches.

The V1 region is primarily viewed as a feed-forward path from the reception of an image at the retina to the higher layers in the visual cortex (Riesenhuber and Poggio, 1999; Serre et al., 2005b). Experiments have shown that this region is responsible for approximately the first 150ms of the human recognition process. Processed results from the V1 region are passed to regions V2 and V4 before being received by the inferotemporal cortex. It is believed that the inferotemporal cortex is responsible for higher levels of recognition such as faces (Desimone et al., 1984, 1985). Modelling this layer takes the form of a combination of learnt salient features from a known dataset coupled with a support vector machine for classification.

In the work of Serre et al. (2005b) a comparison was made between the visual cortex model and the SIFT descriptor (Lowe, 2004) for generalized object recognition using a Support Vector Machine for classification on the Caltech 101 dataset (Fei-Fei et al., 2007). The results demonstrated that the visual cortex based approach outperformed the SIFT method by some margin.

In this work we explore a 3D extension to the visual-cortex model of Mutch and Lowe (2008) that was derived from the work of Serre et al. (2004) which, in turn, followed the visual-cortex standard model (Riesenhuber and Poggio, 2003). The work of Mutch and Lowe (2008) produced excellent recognition rates with small training sets which strengthens its investigation given the limited amount of baggage data available for this work. The model of Mutch and Lowe (2008) comprises a number of steps from input image to output descriptor and will now be discussed in more detail.

Before examining the 3D volumetric extension we will firstly discuss the 2D image-based approach in some detail to establish a sound basis from which to proceed.

7.2 The 2D image-based approach

We follow a simple form of the standard model formulated by Mutch and Lowe (2008) as an extension to the model of Serre et al. (2004). Figure 7.1 (taken directly from Mutch and Lowe, 2008) shows the approach taken by Mutch and Lowe (2008) from input image through to output description.

The first stage is to form a scale-space pyramid comprising N levels ($N = 10$) by re-sampling the input image (using cubic-spline interpolation) with successively smaller scales. At each layer, the image is $2^{1/4}$ smaller (in terms of pixel dimensions) than the last, resulting in a pyramid comprising 2.5 octaves of scale space. Subdi-

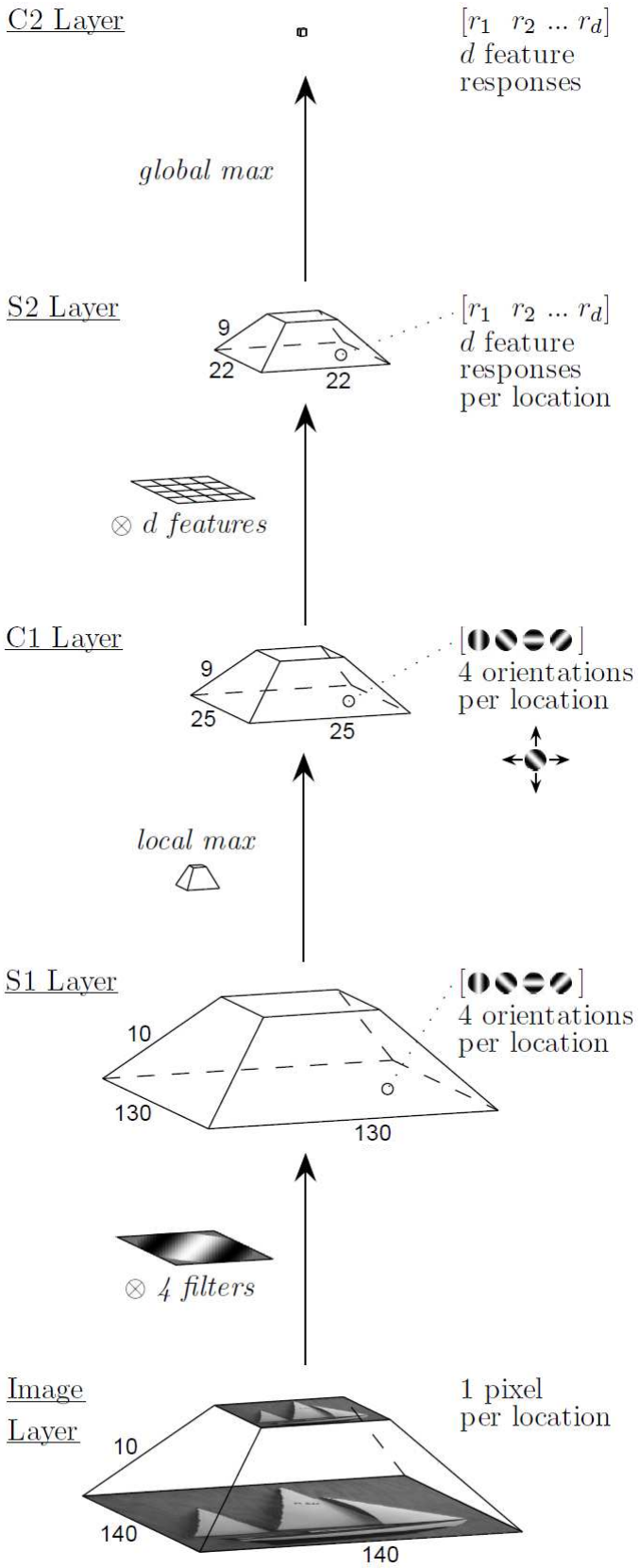


Figure 7.1: 2D flow from input image to output descriptor (taken directly from Mutch and Lowe, 2008)

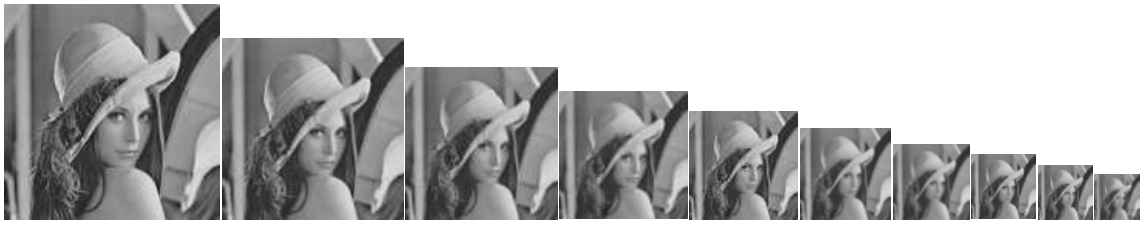


Figure 7.2: Example 2D scale-space pyramid images

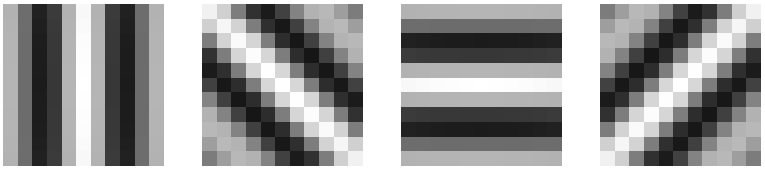


Figure 7.3: 2D Gabor filters: four orientations are used

vision of each octave into 4 layers defines the scale-space resolution. This is similar to the approach taken in earlier work (Chapter 4) where the scale space was divided into 3 layers per octave. This step, introduced by Mutch and Lowe (2008), reduces the computational load in subsequent processing blocks. The original approach of Serre et al. (2005b) kept the image a constant size but varied the size of the filtering kernels in order to detect features at differing scale. The formation of the scale-space pyramid is followed by four processing layers comprising alternate simple (S) and complex (C) functions mirroring the functionality of the human visual cortex. Figure 7.2 shows an example of the scale-space pyramid with a base image 140×140 pixels in size.

The first S layer (S1) is produced by filtering each input pyramid layer using a set of fixed-size Gabor filters of varying orientation. In the 2D case there are four orientations: 0°, 45°, 90°, 135°. Figure 7.3 shows the four filters used in this case, each being 11×11 pixels in size.

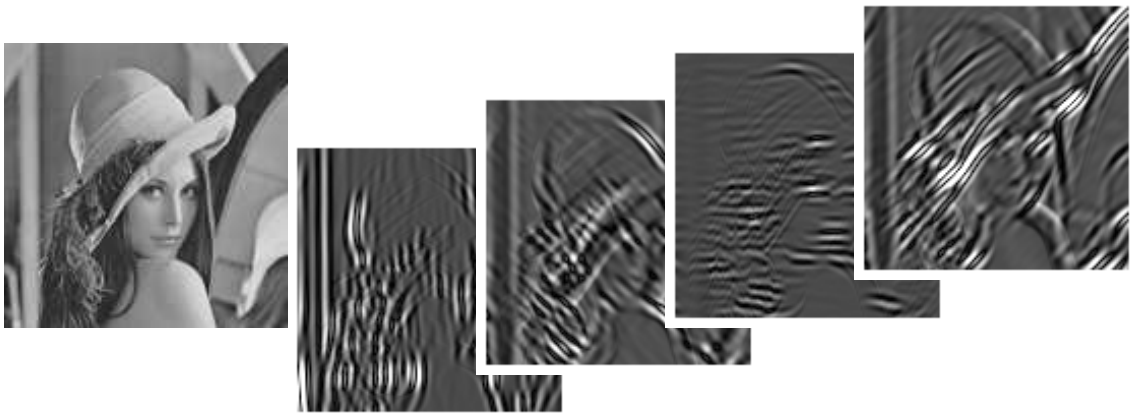
The application of the fixed-size filters to the scale-space pyramid allows features of different size to be extracted: increasingly larger features are revealed as the smaller images in the pyramid are processed. Figure 7.4 shows this for 3 example layers in the pyramid. In Figure 7.4a we see an example image of 140×140 pixels from the base of the scale-space pyramid together with the Gabor-filter responses that result following application of the Gabor filters shown in Figure 7.3. Some of the facial features can be seen in the Gabor responses (lips, nose, eyes). In Figure 7.4b we see an example image from the middle of the scale-space pyramid, in this case comprising 70×70 pixels. We can see the Gabor responses in this case show responses to larger features in the image when compared to the base-image responses (Figure 7.4a). Facial features are no longer shown but instead the responses show

such things as the hat brim, shoulder and background furniture. Finally in Figure 7.4c we see the image at the top of the pyramid (29×29 pixels) from which it can be seen the Gabor-response result from even larger features.

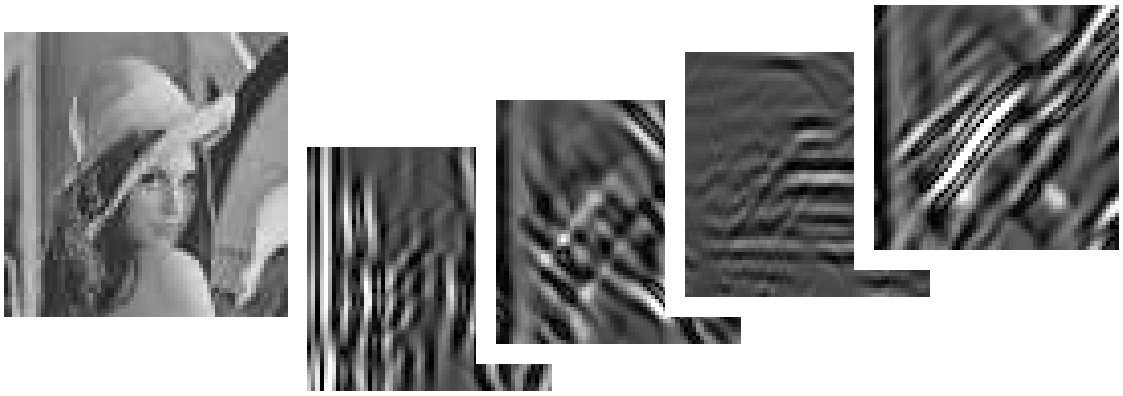
In practice the Gabor-filtering operation on each layer of the pyramid produces vector images: each output pixel contains a vector recording the result of each Gabor-filter response at that location. The definition for the Gabor filters is taken from the work of Serre et al. (2005b) where performance characteristics were taken from V1 parafoveal simple cells and translated to images 140 pixels in height.

Referring to Figure 7.1, taken from the work of Mutch and Lowe (2008), we see that following the S1 layer is the first C layer (C1), produced by applying a localized pooling operation. The pooling operation retains the maximum value for each Gabor filter in the S1 output within a local sub-region of the S1 layer. Similar processes have been observed in neurons within the visual cortex (Hubel and Wiesel, 1959). Figure 7.5 shows a simple example of this operation where we can see four Gabor-filtered images that form one layer in S1. A square pooling window is scanned across each image and the maximum value is stored in the corresponding location in the C1 output image. The pooling window sub-samples the S1 imagery such that the output C1 images are smaller in size. In practice the local pooling function is applied across two scale-space layers and is itself a pyramid in shape so that the same physical area is addressed from each layer. Figure 7.6 shows this operation for one Gabor orientation in the pyramid where we see the adjacent scale-space images and the resultant C1 layer. The pooling operation traverses the S1 image pyramid in both position and scale to produce the C1 layer. As a result of the pooling in scale, there is one fewer C1 pyramid level than the input S1 pyramid. Note that the size of the pooling operation is proportional to the input image size. In Figure 7.6 we see that the pooling area (red square) is smaller in layer $N+1$ than in layer N; this allows the output image to be a size that relates to the size of the layer N image and the degree of sub-sampling that has been applied.

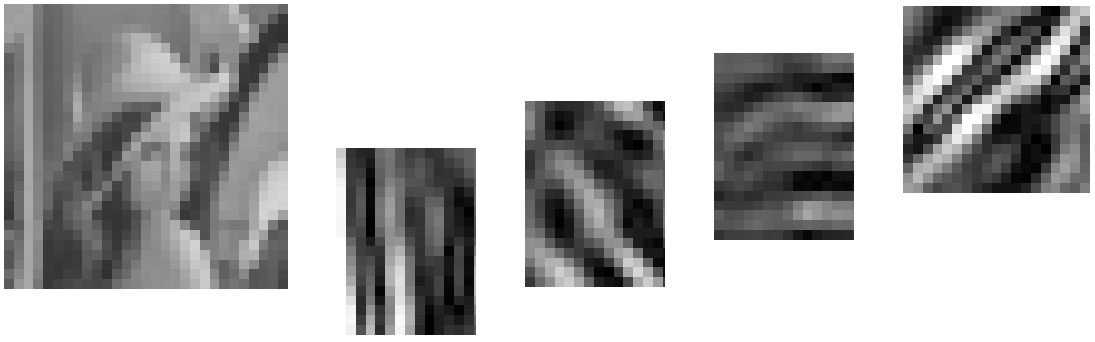
The second S layer (S2) is the start of the modelling of higher level recognition functions in the cortex. It is another filtering stage formed by the comparison of the C1 layer with a set of pre-calculated C1 feature patches that are found to be salient in the recognition task at hand. The selection of the feature patches and their size forms the basis of the recognition system. The formation of the S2 layer in the 3D case is described in Section 7.3.5. The salient feature patches are taken from the C1 layers of a training set of volumes and are chosen to be representative of the object classification task at hand. In the work of Mutch and Lowe (2008) this was achieved by first randomly sampling the training set to provide N_c candidate feature patches.



(a) Base layer: 140×140 pixels



(b) Intermediate layer: 70×70 pixels



(c) Top layer: 29×29 pixels

Figure 7.4: Application of Gabor filters to layers in scale-space pyramid

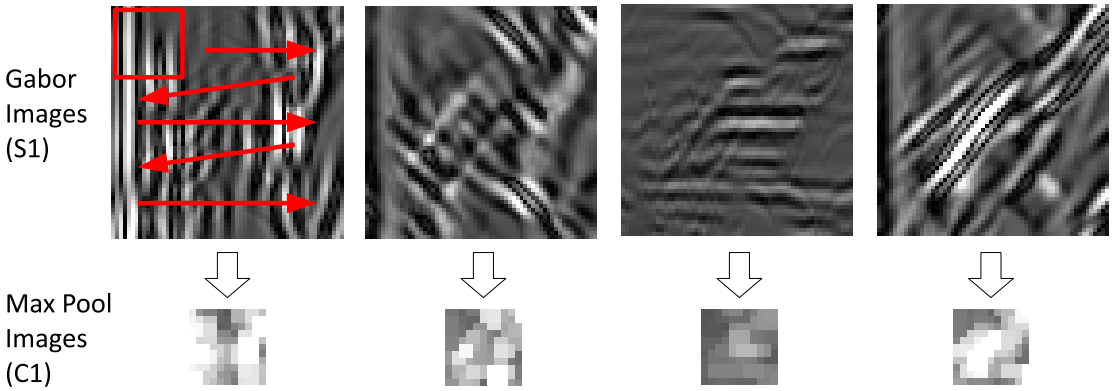


Figure 7.5: Max-pooling operation for one layer: position only

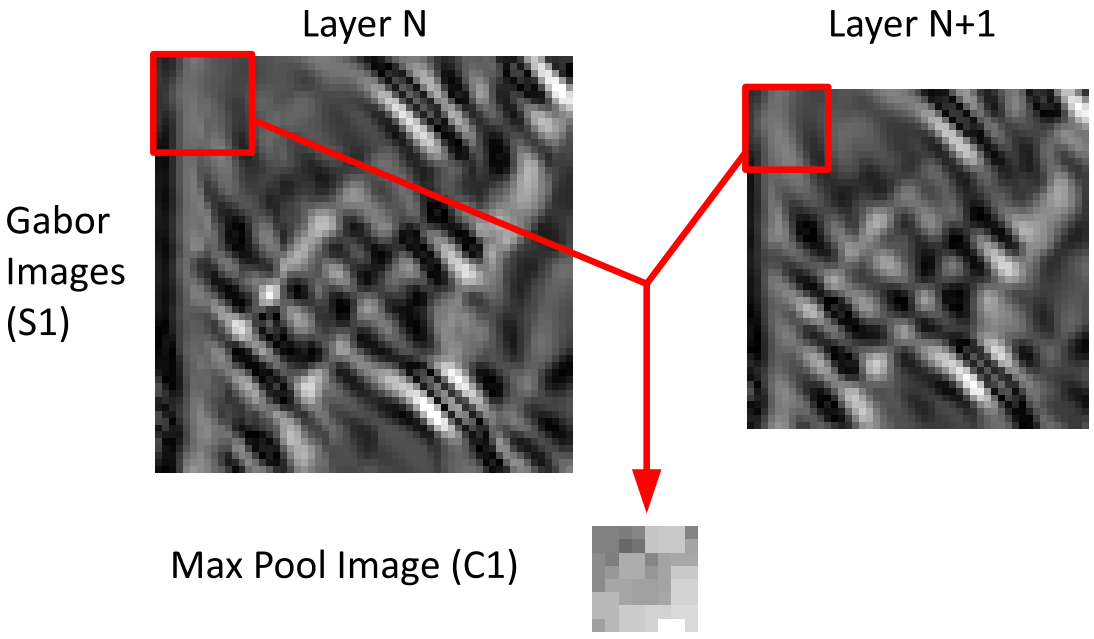


Figure 7.6: Max-pooling in scale space

A support vector machine using a linear kernel was then used to reduce this set to one of N_p feature patches following the work of Mladenić et al. (2004).

The final layer, C2, is formed by taking the maximum value of the S2 layer for each of the salient feature patches in order to form a descriptor vector and is described in Section 7.3.6. This descriptor vector is then used as a training vector for supervised training of an SVM or as input to the SVM for classification in determining recognition performance.

It has been noted that the model is inherently not invariant to rotation (Serre et al., 2004). Rotational invariance is a critical aspect of our work so it is essential that the training data contains examples of target items in a variety of orientations. It would be possible to increase the amount of data by using simple image transformations (rotation, reflections), as performed by Mutch and Lowe (2008), but this has not been implemented at present. Some care would be needed if this were done as the imaging artefacts (streaks, shadows) radiate in the xy plane (see Section 3.1.3); an arbitrary transformation would not maintain this aspect of the imagery.

7.3 Extension to 3D

Extending the work of Mutch and Lowe (2008) from 2D recognition to 3D is conceptually straightforward but computationally more intensive. The image pyramid becomes a volumetric-image pyramid and extending the Gabor-filtering stage of S1 turns 2D filters into 3D volumetric filters. Subsequent image patches become volumetric patches. We will now describe the extension of the visual-cortex model into 3D in detail.

7.3.1 Formation of volumetric scale-space pyramid

Following on from Mutch and Lowe (2008) we form a volumetric scale-space pyramid comprising 10 levels ($L = 0, 1, \ldots, 9$). Each level is $2^{1/4}$ smaller than the last, that is, the number of voxels in each dimension reduces from the base layer upwards by a factor of ≈ 0.8409. Bi-cubic interpolation is used to produce each layer from the previous, with care being taken regarding the definition of the origin of each volume. We can view the volumes in a pyramid according to the number of voxels in each dimension (Figure 7.7) but it is useful to note that we can also regard each volume as being the same physical size (in cm) with the size of voxels increasing as the pyramid is built. Interpreting the volumes in this manner is useful during the generation of the C1 layer (Section 7.3.4) where we must ensure identical absolute location of points between layers.

Prior work in 2D imagery uses the scale-space pyramid to allow the same object to be recognized even when suffering from perspective distortion (i.e. objects closer to the camera appear larger than those in the distance). In the work of Serre et al. (2005b) base images were rescaled to a width of 140 pixels whilst for Mutch and Lowe (2008) it was the height that was rescaled to 140 pixels prior to construction of the scale-space pyramid. This ensured that objects within the images were observed at a similar pixel resolution and allowed the scale-space pyramid to cover just 2.5 octaves. Without the rescaling process it is conceivable that the scale-space pyramid would have been required to cover many more octaves. In our case the CT imagery relates objects to a real physical dimension (in cm) and does not suffer from perspective distortion, so we do not require a rescaling process.

7.3.2 3D Gabor filters

Gabor filters are simple edge detectors that combine a sinusoidal response with a Gaussian envelope. The size of edge feature that can be detected is determined by the wavelength, λ, of the sinusoid and the effective width, σ, of the response. When extended to more than one dimension an additional parameter is introduced: the aspect ratio, γ. This term limits the response in directions orthogonal to that of the main sinusoid.

We extend the 2D Gabor definition from Mutch and Lowe (2008) into 3D using the directions given from the 20 vertices of a dodecahedron. The vertices are in pairs on opposite sides of the dodecahedron resulting in 10 unique directions and hence 10 Gabor filters. The vertices are defined as coordinates using the golden ratio:

$$\Phi = \frac{1+\sqrt{5}}{2} \tag{7.1}$$

$$\psi = 1/\Phi \tag{7.2}$$

We define the 10 direction vectors as follows:

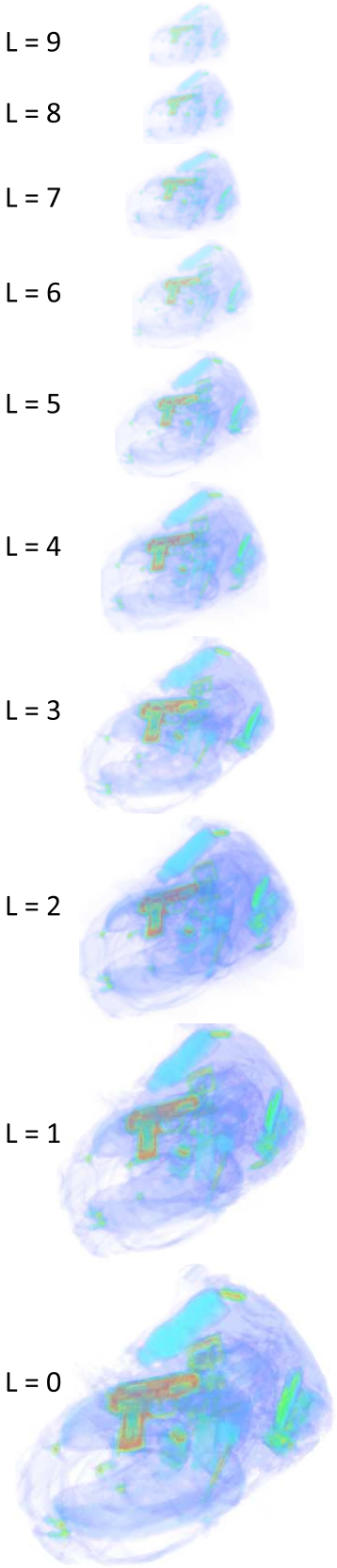


Figure 7.7: Volumetric pyramid scale-space example

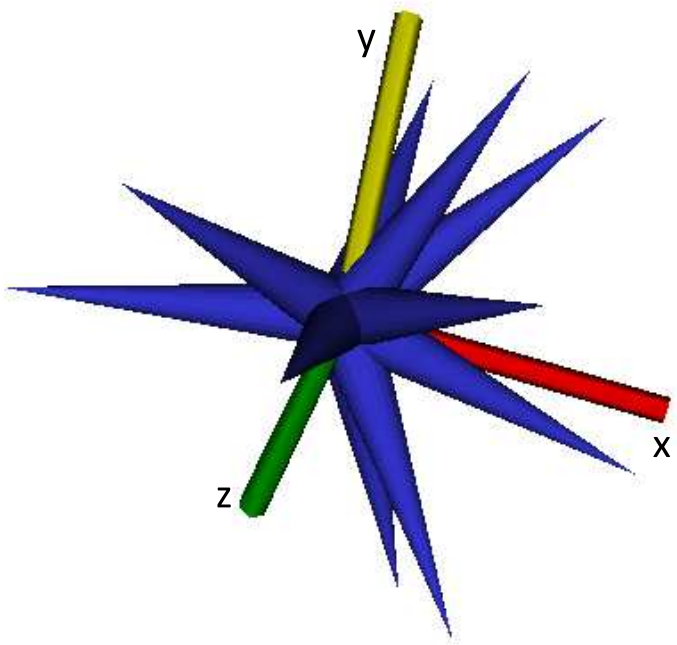


Figure 7.8: Directions formed from dodecahedron vertices

$$\begin{bmatrix} x_v \\ y_v \\ z_v \end{bmatrix} = \begin{bmatrix} 0 \\ \psi \\ \Phi \end{bmatrix}, \begin{bmatrix} \Phi \\ 0 \\ \psi \end{bmatrix}, \begin{bmatrix} \psi \\ \Phi \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ \psi \\ -\Phi \end{bmatrix}, \begin{bmatrix} -\Phi \\ 0 \\ \psi \end{bmatrix}, \begin{bmatrix} \psi \\ -\Phi \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} \tag{7.3}$$

Figure 7.8 shows the directions in 3D space formed from the dodecahedron coordinates.

We convert to polar coordinates by defining the azimuth, θ, and elevation, ϕ, as follows:

$$\theta = \arctan\left(y_v / x_v\right) \tag{7.4}$$

$$\phi = \arctan\left(\frac{z_v}{\sqrt{x_v^2 + y_v^2}}\right) \tag{7.5}$$

Note that, in practice, we resolve θ into the range $[-\pi, \pi]$ as we can identify the correct quadrant from (x_v, y_v). From these definitions we create two matrices that

specify rotations around the y and z axes:

$$R_y = \begin{bmatrix} \cos\phi & 0 & \sin\phi \\ 0 & 1 & 0 \\ -\sin\phi & 0 & \cos\phi \end{bmatrix} \tag{7.6}$$

$$R_z = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{7.7}$$

We can now define a Gabor filter in 3D for a given voxel at location $\begin{bmatrix} x & y & z \end{bmatrix}^T$. Using the rotation matrices, R_y and R_z, we form a new coordinate set:

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{bmatrix} = R_y R_z \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{7.8}$$

From which the Gabor filter is defined as:

$$G(\hat{x}, \hat{y}, \hat{z}) = \exp\left[-\frac{1}{2\sigma^2}\left(\hat{x}^2 + \gamma^2\hat{y}^2 + \gamma^2\hat{z}^2\right)\right]\cos\left(\frac{2\pi\hat{x}}{\lambda}\right), \tag{7.9}$$

where γ is the aspect ratio, σ is the effective width and λ is the wavelength. Following Mutch and Lowe (2008) we define the size of each Gabor filter as an $N_G \times N_G \times N_G$ voxel volume with $N_G = 11$ where x and y vary between -5 and $+5$. We further set $\gamma = 0.3$, $\sigma = 4.5$ and $\lambda = 5.6$ as defined by Serre et al. (2005b) and followed by Mutch and Lowe (2008). As each filter is heavily truncated we further adjust each to have zero mean and then normalize to give a unity sum of squares. Given that the CT imagery has real-world dimensions we can relate the value chosen for the wavelength to real-world features. Given $\lambda = 5.6$ in *voxels* implies that at the base level ($L = 0$), where each voxel is a $0.25cm$ cube, we have $\lambda_0 = 1.4cm$. At the top of the pyramid ($L = 9$), where each voxel is a $1.41cm$ cube, we have $\lambda_9 = 7.9cm$. These values for λ indicate the range in the size of the features that are being extracted by the Gabor filtering.

Figure 7.9 shows the 10 Gabor filters displayed as 3D volumes where we can see the varying orientations and truncated extent.

7.3.3 S1 layer

The S1 layer is formed through application of the Gabor filters to each volume of the volume pyramid following the method described in Mutch and Lowe (2008).

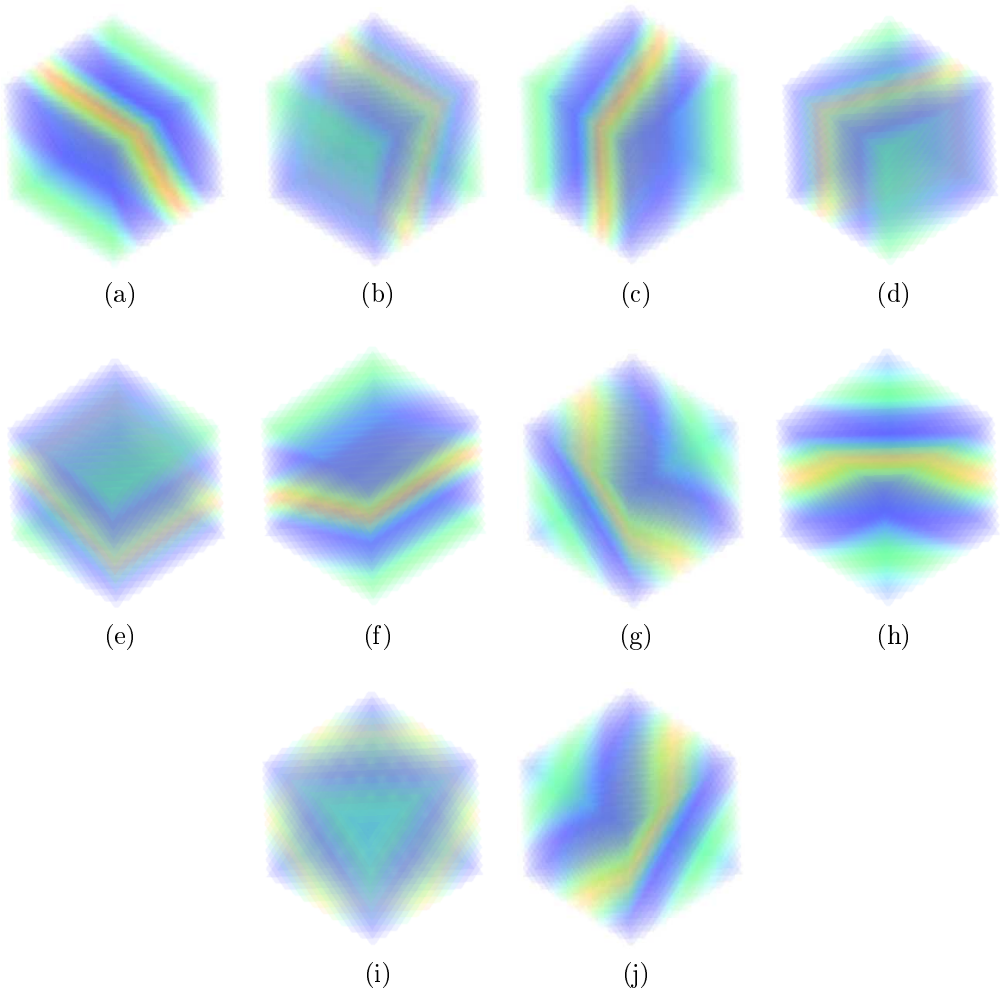


Figure 7.9: Extended 3D Gabor filters used in the S1 layer

The response to a given volumetric patch of voxels, X, in each volume to a Gabor filter, G, is defined by:

$$R(X,\, G) = \left| \frac{\sum_i X_i G_i}{\sqrt{X_i^2}} \right| \tag{7.10}$$

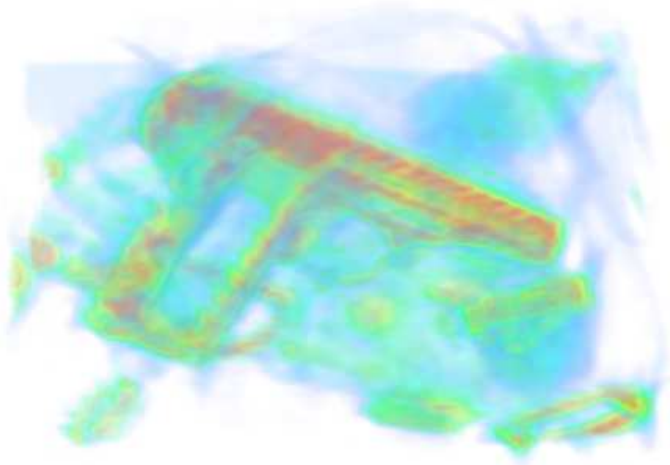
This response is motivated by observation of neurons within V1 (Hubel and Wiesel, 1968) where the response to a given edge relates to its orientation and intensity. We do not extend this filtering step outside the volume extent, so the output response volume will be smaller than the input by $N_G - 1$ voxels.

Examples of the Gabor-filter responses are given in Figure 7.10 and Figure 7.11. In Figure 7.10 we see the responses at the base layer of scale space ($L = 0$). Figure 7.10a shows the base-layer volume containing a pistol and several items of clutter (golf balls, belt buckle, etc.). Figure 7.10b shows the Gabor-filter responses for this input volume where we can see that the response of each filter reflects the spatial arrangement of objects in the space. As we move up the scale-space pyramid the Gabor-filter responses appear more and more coarse, reflecting the larger size of features detected. Figure 7.11a shows the same pistol volume re-sampled further up the pyramid ($L = 4$) with Figure 7.11b showing the associated Gabor-filter responses. We can see from this that the responses are larger in size when compared to those in Figure 7.10b.

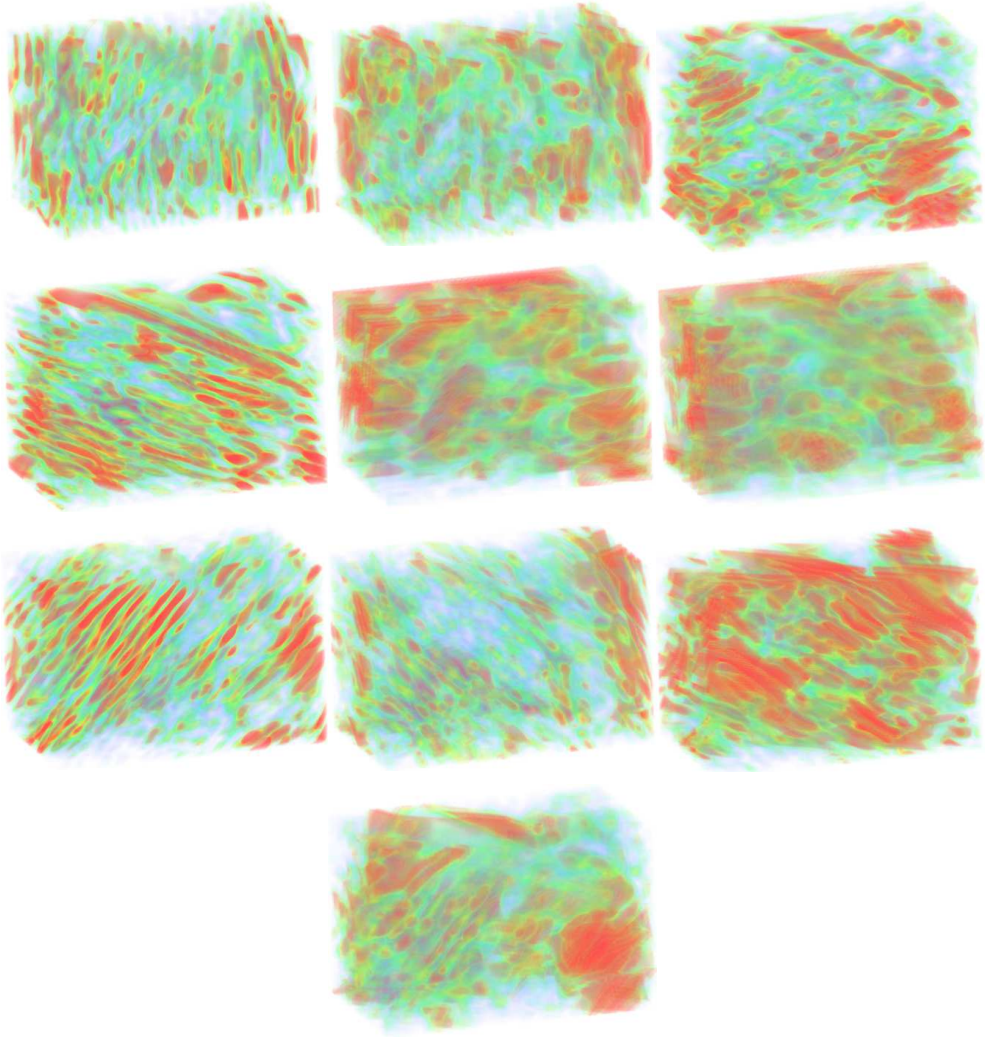
Note that, in practice, the Gabor-filtered output volumes are combined to form a single vector volume, in which each voxel contains a vector of 10 elements representing the response to each of the 10 Gabor filters. This allows a simpler model to be implemented.

7.3.4 C1 layer

The functionality of this layer is to provide local invariance by retaining peak S1 responses within localized regions as a means to mimic the functionality of the complex cells in V1. Extending the work of Mutch and Lowe (2008), a volumetric pyramid volume comprising 2 scales with $10 \times 10 \times 10$ voxels at the lowest scale is scanned through the input S1 data recording the peak S1 response for each Gabor orientation within the S1 data. Sub-sampling of the data takes place by adjusting the max-pooling filter location in steps of 5 voxels. The max-pooling filter has nominally $8.4 \times 8.4 \times 8.4$ voxels in its higher layer. When applying this we round to the nearest voxel when deciding if a voxel in the higher image layer can contribute

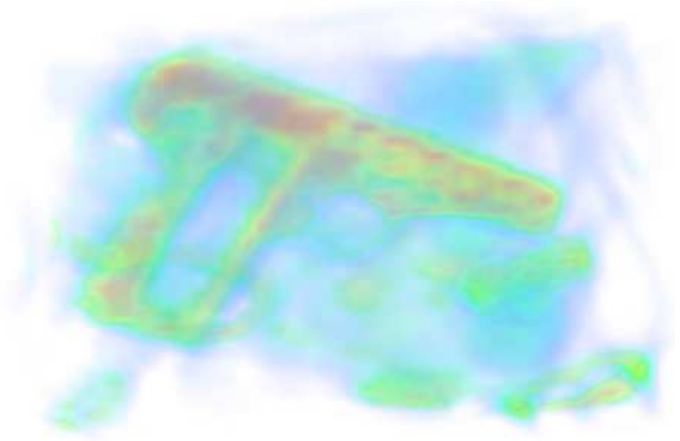


(a) Input volume: pistol with clutter

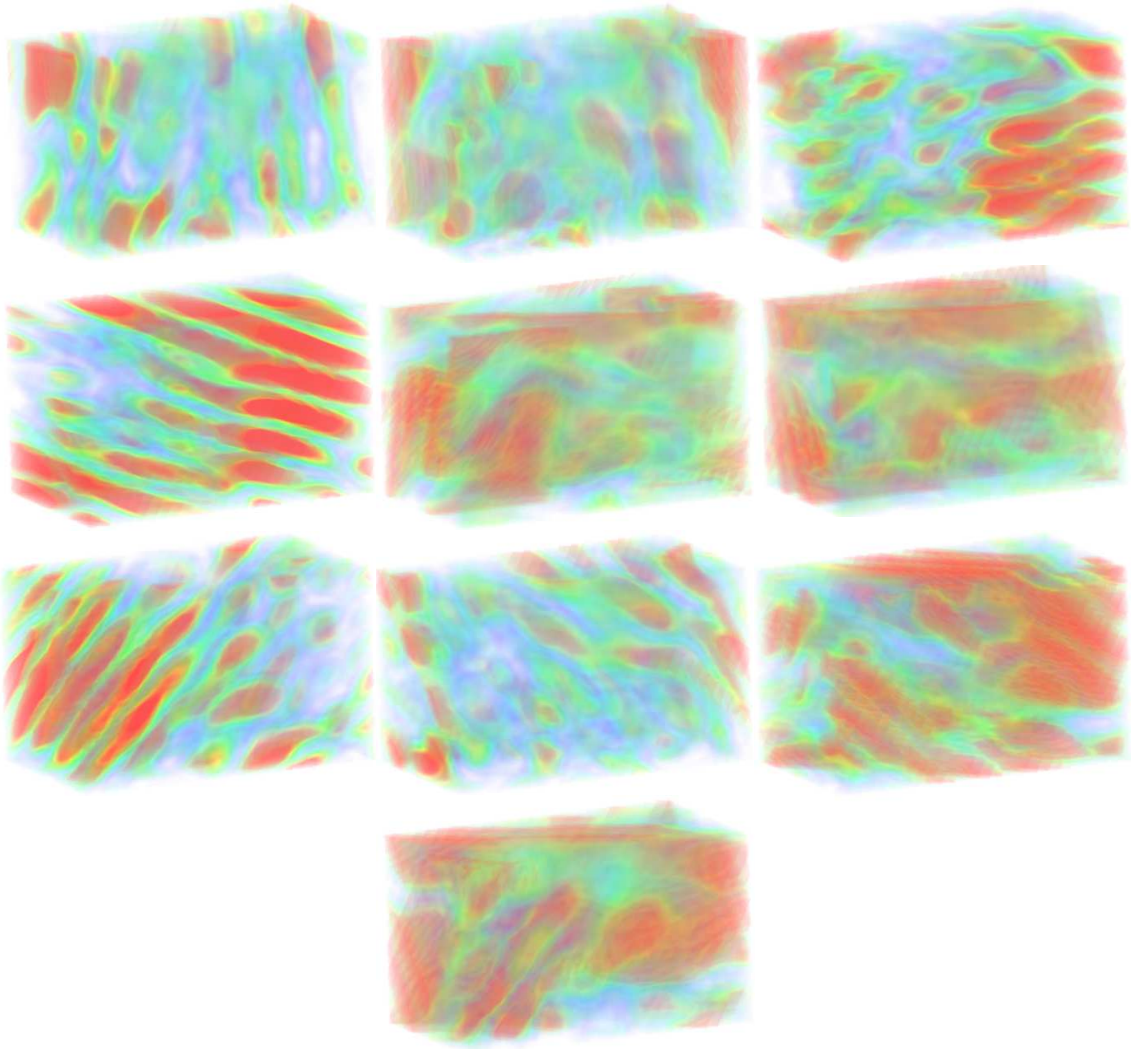


(b) Response for each Gabor filter

Figure 7.10: Example response to Gabor filter at level 0 of scale-space pyramid



(a) Input volume: pistol with clutter



(b) Response for each Gabor filter

Figure 7.11: Example response to Gabor filter at level 4 of scale-space pyramid

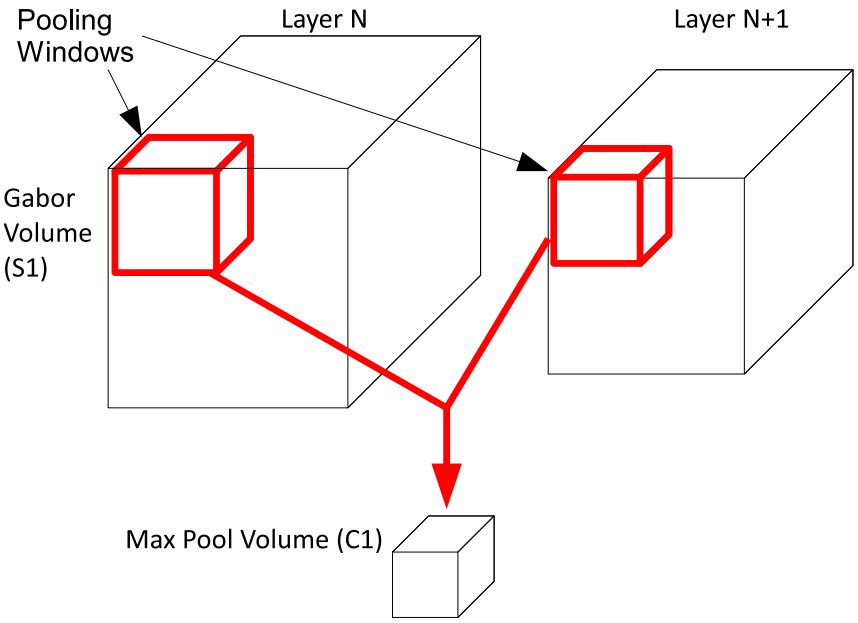


Figure 7.12: Max pooling in position and scale in 3D

to the max-pooling function at any given central location.

Figure 7.12 illustrates this process in 3D where we see the pooling window being applied in two volumes from the S1 scale-space pyramid resulting in the C1-output volume. Note that the input volumes contain vectors at each voxel recording the responses from each Gabor filter and consequently the output C1 volume also contains vectors of the same dimension. It is worth remembering that the filter responses are considered as separate entities: we record the maximum response for each Gabor filter in the pooling window.

The result of the C1 layer process is again a pyramid structure comprising $N_s - 1$ scales with smaller volume dimensions which result from the max-pooling volume sub-sampling.

7.3.5 S2 layer

The S2 layer executes template matching between the C1 layer and a set of predetermined patches (see Section 7.4). This stage represents the beginning of a higher level of recognition in the visual cortex located in V4/inferotemporal cortex.

Following Mutch and Lowe (2008), the response, $R(X, P)$, of a patch of C1 units, X, to a predetermined feature, P, is given by a radial basis function:

$$R(X,\ P) = \exp\left(-\frac{\|X - P\|^2}{2\sigma^2\alpha}\right) \tag{7.11}$$

In our work, both X and P are $n \times n \times n$ voxels in size with each voxel containing

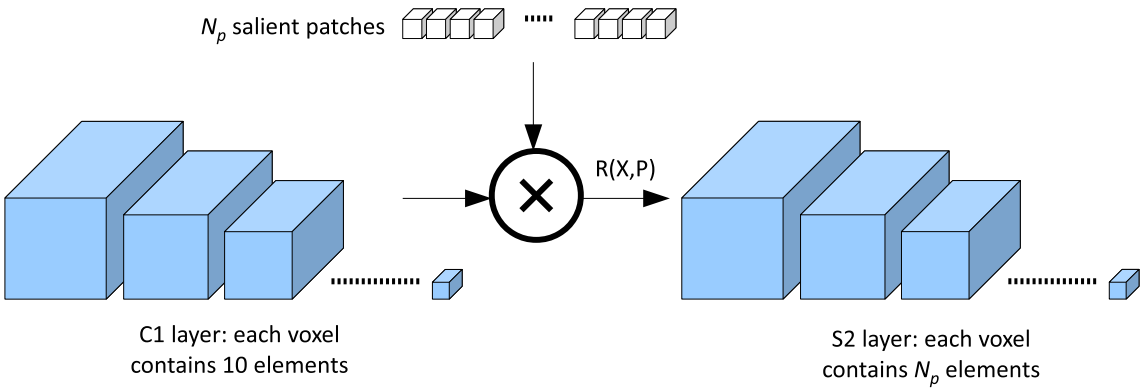


Figure 7.13: Formation of S2 layer

a vector of 10 values derived from the 10 Gabor filters used in the creation of the S1 pyramid layer (Section 7.3.3). We follow both Mutch and Lowe (2008) and Serre et al. (2004) in setting $\sigma = 1.0$. The setting of α is used by Mutch and Lowe (2008) to provide a normalization term for patches of differing size. For the 2D case, Mutch and Lowe (2008) had patches of size $n \times n$ where $n = \{4, 8, 12, 16\}$. The normalization term for the 2D case is then $\alpha = (n/4)^2$ such that α normalizes the patch response, R, relative to the smallest patch size dimension being used, in their case 4. For 3D we modify this term to reflect the increased dimension, again assuming a lower setting of $n = 4$:

$$\alpha = \left(\frac{n}{4}\right)^3 \tag{7.12}$$

For example, using $n = 4$ will result in a 3D patch comprising 640 elements.

The response (Equation (7.11)) for each salient patch is calculated at every location in the C1 scale-space pyramid so that the S2 output is another scale-space volumetric pyramid, but in this case each voxel contains a vector of N_p response values. This is illustrated in Figure 7.13 where we see the evaluation of the response between the C1 layer and salient patches resulting in the S2 layer as output. It is worth noting that this process will reduce the size of each volume in the S2 layer when compared to the C1 input due to the size of the salient patch being used.

7.3.6 C2 layer

This layer forms the bag-of-features type vector (Sivic and Zisserman, 2003; Csurka et al., 2004) for presentation to a support vector machine and is straightforward in nature. We establish a feature-response vector by taking the largest patch response for each feature in the S2 layer of the baggage item being analyzed (Figure 7.14). For example, the first element in the feature-response vector is obtained by examining

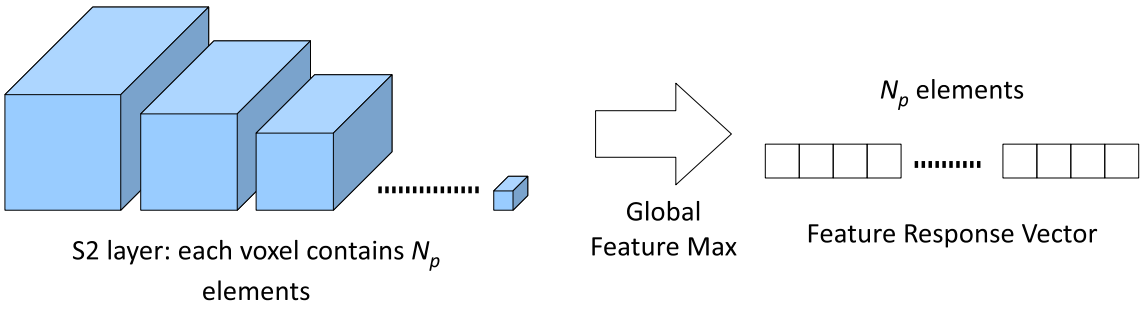


Figure 7.14: Formation of C2 layer response vector

the first element in each voxel of the S2 scale-space pyramid and retaining the largest value. This is repeated for each element so that, given N_p salient feature patches, we now have a vector of N_p values that describes the volumetric imagery in terms that can be used by a machine-learning algorithm for training or classification.

7.4 Feature selection

The choice of salient patches for classification from the C1 layer of a training set of volumes is a key aspect in the recognition methodology. We wish to use patches that make a strong contribution to the classification of a given volume as either a "threat" or as "no threat".

We follow the work of Mutch and Lowe (2008) by first randomly choosing N_r ($N_r = 12,000$) patches from the C1 layers of a training set of volumes. We choose patches of size $4 \times 4 \times 4$ voxels, each of which contain 10 Gabor-orientation results. Using a linear SVM we can select the patches that contribute most to the classification process in order to remove patches that do not significantly contribute to the solution. We aim to reduce the number of patches used for classification from N_r ($N_r = 12,000$) to N_p ($N_p = 1500$).

The work of Mladenić et al. (2004) proposed the use of linear support vector machines in the identification of salient features for classification tasks. The SVM derives a hyperplane whose normal can indicate relative strengths of candidate features in the classification task (see Section 6.4). This approach was used by Mutch and Lowe (2008) and we choose to follow that method. By way of explanation, consider Figure 7.15 which illustrates a simple class-separation task in three dimensions. Two classes (A, B) can be separated by a linear plane (P). In this example the normal to the plane has 3 components: $[\eta_x\ \eta_y\ \eta_z]$. It can be seen that the largest component of the normal is η_y followed by η_x then η_z. This is indicative of the fact that the classification of a point into class A or B is most dependent on its y

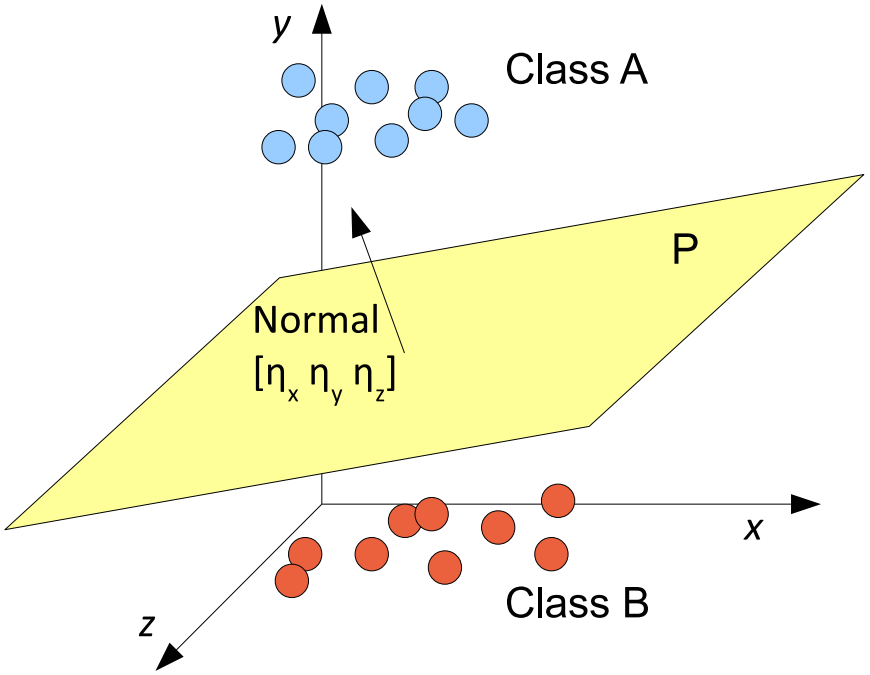


Figure 7.15: Example of class separation using a plane

component followed by its x component then its z component. In the classification process within the SVM, each axes represents the response to a particular feature. There will be therefore many more dimensions than this simple example. However, the principle of using the normal components to rank feature significance has been proven to work successfully (Mladenić et al., 2004; Mutch and Lowe, 2008).

Figure 7.16 shows how the selection process follows a pyramid structure in order to filter out poor quality patches and leave N_p ($N_p = 1500$) patches as the bag of visual words. The initial set of N_r patches are split into four groups of N_p patches ($N_p = 3000$) and these form a candidate set of patches for the matching process. In each case a linear SVM is trained using the candidate patches and the training volumes. The trained SVM separates the training data using a hyperplane. Given ν support vectors, $x_{v,i}$, each with N_p components ($i = 1..N_p$), each component of the normal to the hyperplane η_i is given by:

$$\eta_i = \sum_v \alpha_v x_{v,i}$$

where α_v define the hyperplane and are calculated as part of the SVM algorithm (Cortes and Vapnik, 1995).

If we sort the normal components by magnitude we can estimate the contribution that each candidate patch has made to the classification process. Patches that have a relatively high normal component magnitude, $|\eta_i|$, have a greater influence on

the classification result and so we wish to retain those as salient features (Mladenić et al., 2004). Following sorting, we reject the least influential half of the patches.

Figure 7.16 illustrates the hierarchical arrangement that reduces the initial set of randomly selected C1 layer patches down to the final set of salient patches. The first selection layer reduces the number of patches from 12000 (in 4 groups of 3000) down to 6000 (in 2 groups of 3000). These patches are then reduced to 3000 before the final selection layer that selects the top 1500 patches to be used for the classification process. For each set of 3000 patches a new SVM is trained and its normals examined to derive the reduced set of salient patches.

7.5 Machine-learning methodology

Derivation of the classification patches (Section 7.4) allows the C2 layer (Section 7.3.6) to be formed. The C2 layer comprises a vector of response values that describe the baggage item and this can be used in training and testing of the classification methodology. We follow our approach from Chapter 6 and use a support vector machine for the classification task.

Figure 7.17 shows the training and testing methodology where we see both training and test sets of C2 layer response vectors as input to a support vector machine. As before the SVM parameters are chosen through a grid search using a ten-fold cross validation on the training set. The parameters which result in the lowest number of misclassifications are chosen and used to retrain the SVM using the complete training data. Once trained, the SVM is presented with the test set and the classification performance (true-positive, false-positive rates) examined.

7.6 Results

We now present the results obtained in the classification of handgun and bottle datasets. It is important to note that, due to the limited amount of data available, it was not possible to derive three separate datasets: handguns, bottles, clear. Instead the data is partitioned such that bottles are considered as part of the clear set for the handgun classification test and vice versa. As for the analysis of the codebook approach (Chapter 6), we use true-positive rate, false-positive rate, precision and recall to assess performance (see Appendix C).

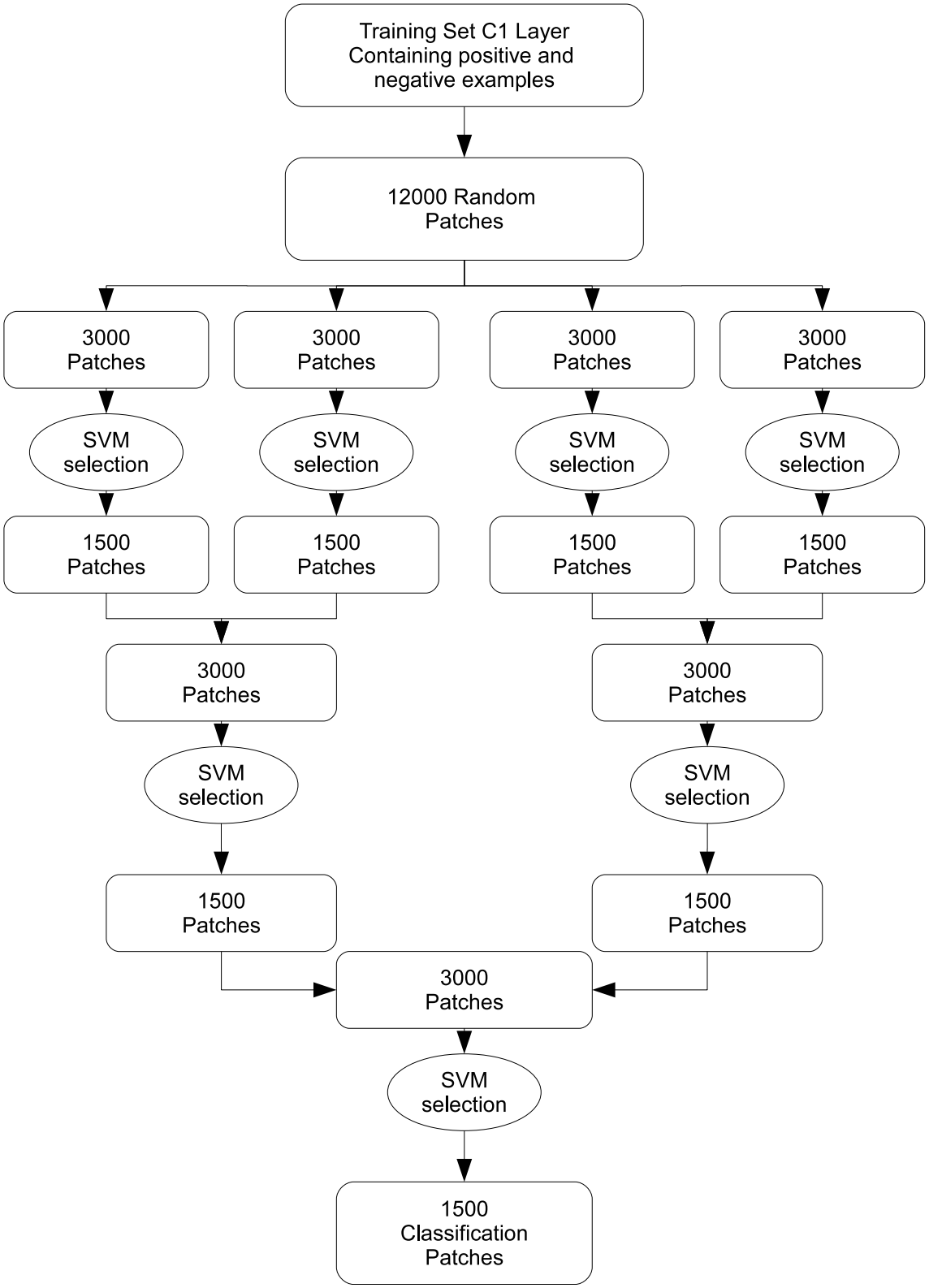


Figure 7.16: Selection of classification patches from an initial random set is achieved using a pyramid of linear SVM selection functions.

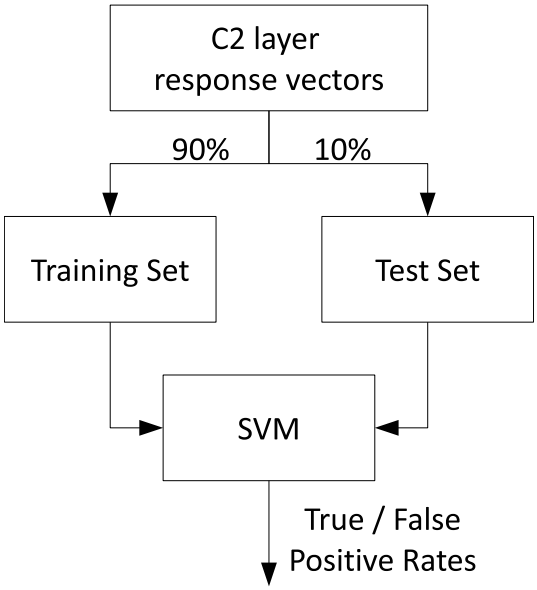


Figure 7.17: SVM usage for classification

Item	Quantity
Threat	284
Clear	971

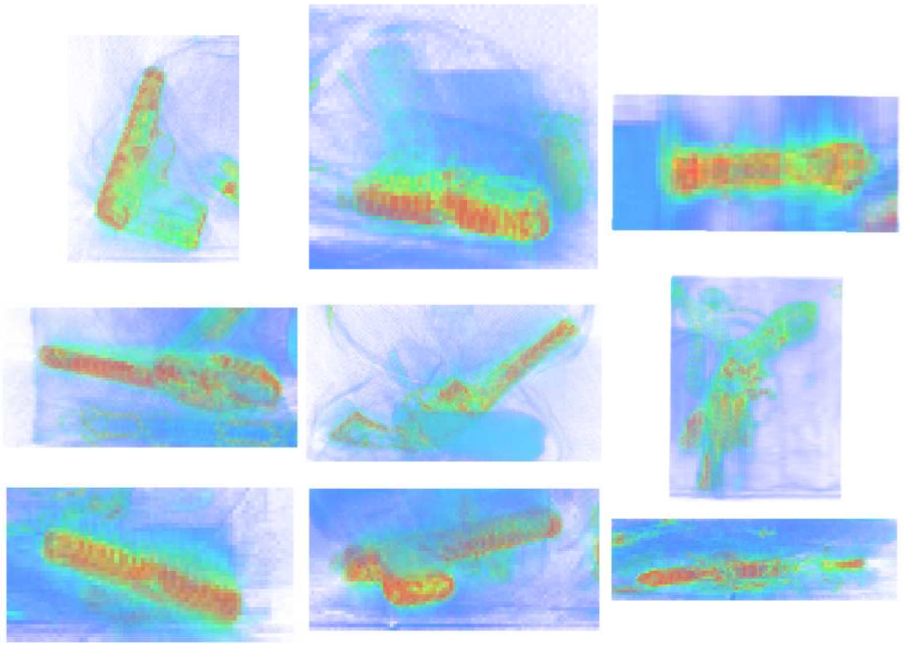
Table 7.1: Handgun group dataset

7.6.1 Handgun results

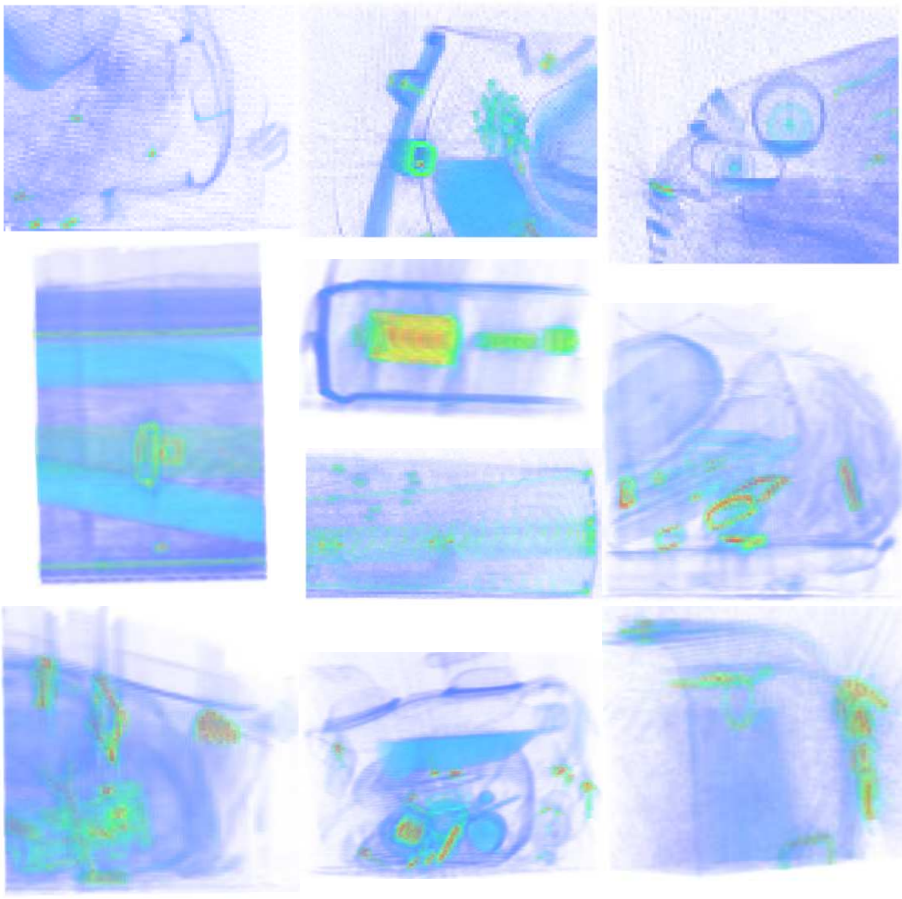
A ten-fold cross validation was performed on the same handgun sub-volume dataset used in Chapter 5. Table 7.1 shows the number of threat and non-threat items in this dataset. Figure 7.18 shows some examples of the threat and clear datasets. Figure 7.18a shows some of pistols and revolvers used in the experiment where the variety of orientations can be seen. Figure 7.18b shows some of the clutter present in the clear volumes including bottles, batteries, clothing and electronic circuitry.

Table 7.2 shows the results for each fold where we can see excellent performance in both true-positive rate and false-positive rate. An overall true-positive rate of 96.8% $\pm$ 2.6% is recorded with a low false-positive rate of 1.1% $\pm$ 0.9%. We also calculate the precision and recall and we can again see good performance with a precision of 0.962 and recall of 0.968.

We are interested in the cases where a misclassification has occurred as these could indicate the nature of the software model that needs altering. Figure 7.19 shows the nine handgun sub-volumes that were incorrectly classified as "clear" during this test. There is no obvious reason for this misclassification: a variety of gun types (pistols and revolvers) and models are represented with a diverse set of orientations.



(a) Example handguns



(b) Example clear

Figure 7.18: Example volumes used for handgun experiment

Fold	True-positive rate (%)	False-positive rate (%)	Precision	Recall
0	96.4 (27/28)	1.0 (1/97)	0.964	0.964
1	92.9 (26/28)	0.0 (0/97)	1.000	0.929
2	92.9 (26/28)	1.0 (1/97)	0.963	0.929
3	96.4 (27/28)	3.1 (3/97)	0.900	0.964
4	96.4 (27/28)	0.0 (0/97)	1.000	0.964
5	100.0 (28/28)	1.0 (1/97)	0.966	1.000
6	96.4 (27/28)	2.1 (2/97)	0.931	0.964
7	100.0 (28/28)	1.0 (1/97)	0.966	1.000
8	96.4 (27/28)	1.0 (1/97)	0.964	0.964
9	100.0 (32/32)	1.0 (1/98)	0.970	1.000
Mean	96.8 ± 2.6 (275/284)	1.1 ± 0.9 (11/971)	0.962 ± 0.029	0.968 ± 0.026

Table 7.2: Handgun sub-volume fold results

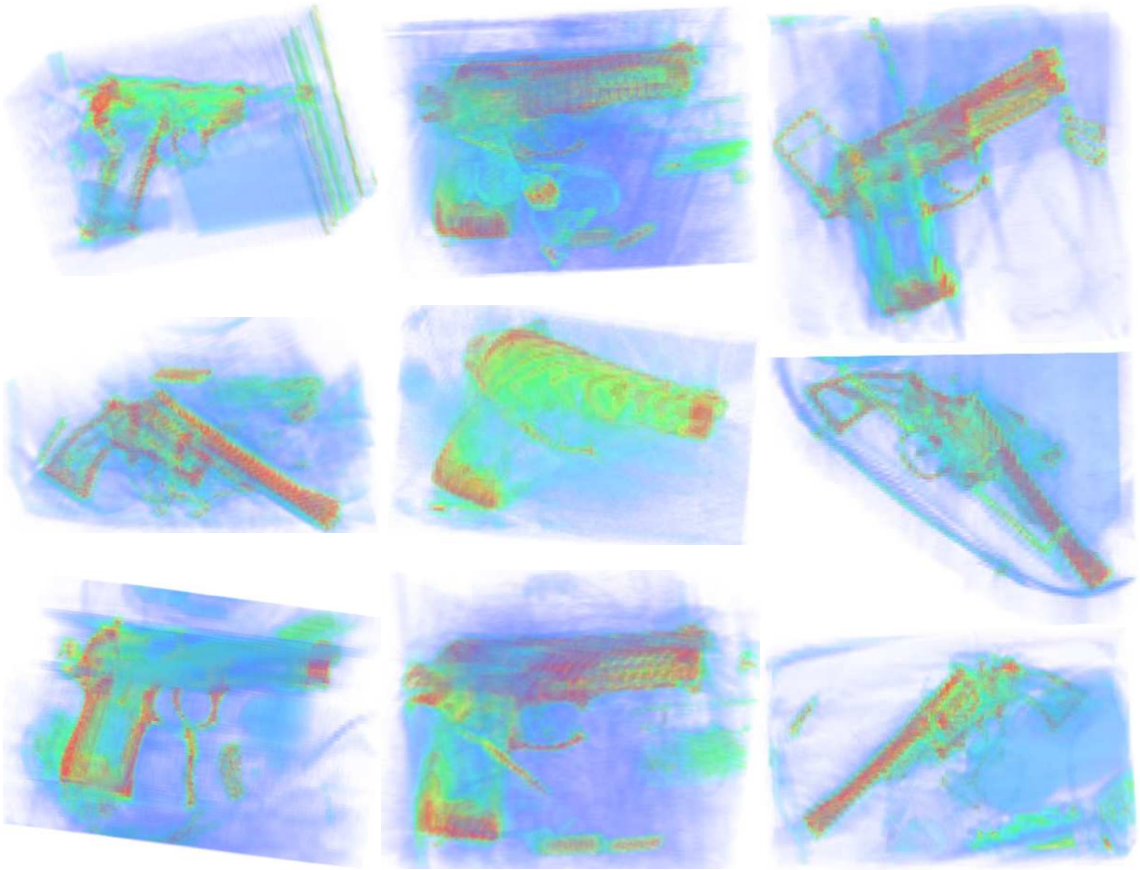


Figure 7.19: Incorrectly classified handguns

Table 7.3: Liquid Container Group Datasets

Item	Quantity
Threat	534
Clear	1170

Figure 7.20 shows the eleven clear sub-volumes that were incorrectly classified as containing a handgun. Again there is no obvious attributable phenomenon behind these misclassifications. One volume does contain a selection of batteries which we could imagine might be confused for a gun part though there are plenty of such volumes in the dataset and only one is misclassified.

Overall we can see from the examples shown in Figure 7.19, Figure 7.20 and Table 7.2 that over a robust cross-validation methodology the technique performs with very high precision and recall statistics matched by a mean 96.8% true-positive rate and mean 1.1% false-positive rate. This is performed over a realistic test set of baggage imagery as laid out in Section 6.4.1.

7.6.2 Liquid container results

We follow the same process for bottles as for handguns. A ten-fold cross validation was performed on the same bottle sub-volume dataset used in Chapter 5. Table 7.3 shows the number of threat and non-threat items in this dataset. Figure 7.21 shows some examples of the threat and clear datasets where Figure 7.21a illustrates the variety of bottles and Figure 7.21b shows some of the clutter present in the clear volumes for this case.

Table 7.4 shows the results for each fold where we can see excellent performance in both true-positive rate and false-positive rate. An overall true-positive rate of 96.6% $\pm$ 3.2% is recorded with a low false-positive rate of 1.0% $\pm$ 1.6%. We again calculate the precision and recall and we see good performance with a precision of 0.977 and recall of 0.966.

Figure 7.22 shows the sub-volumes containing containers that were incorrectly classified as "clear" during this test. It is unclear as to the reason for misclassification for most of these examples. However, some of the examples (marked with '*') contain only partial bottles (an error in the formation of the dataset) which could easily be misclassified as clear given that the bottle part is close to the volume edge and may not result in a feature being generated in the S1/C1 phase of the formulation.

Figure 7.23 shows the clear sub-volumes that were incorrectly classified as bottle. In some instances we can see a possible explanation. We can see that a book may be misclassified as a bottle: it is a similar size to a bottle and has a similar density

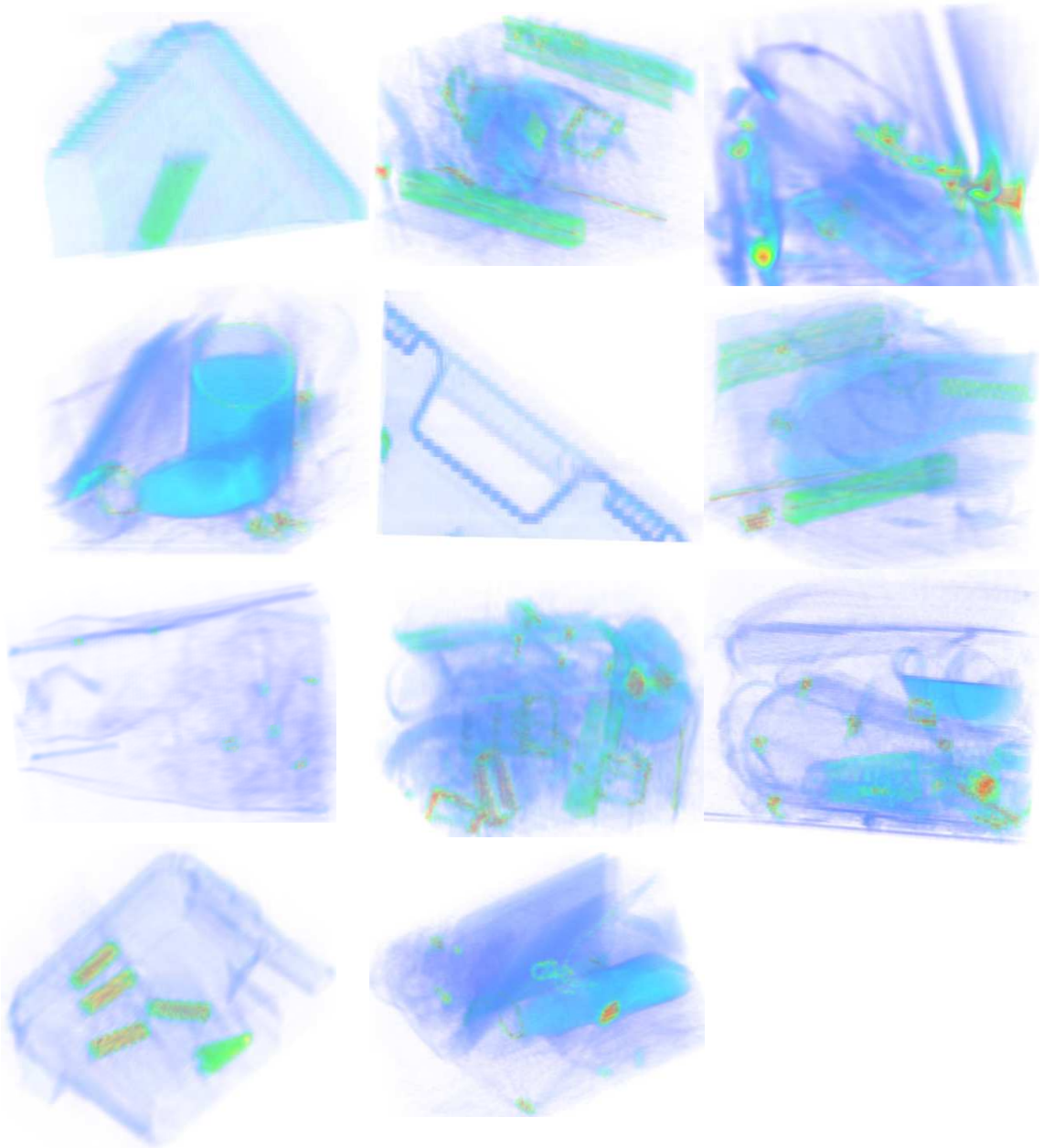
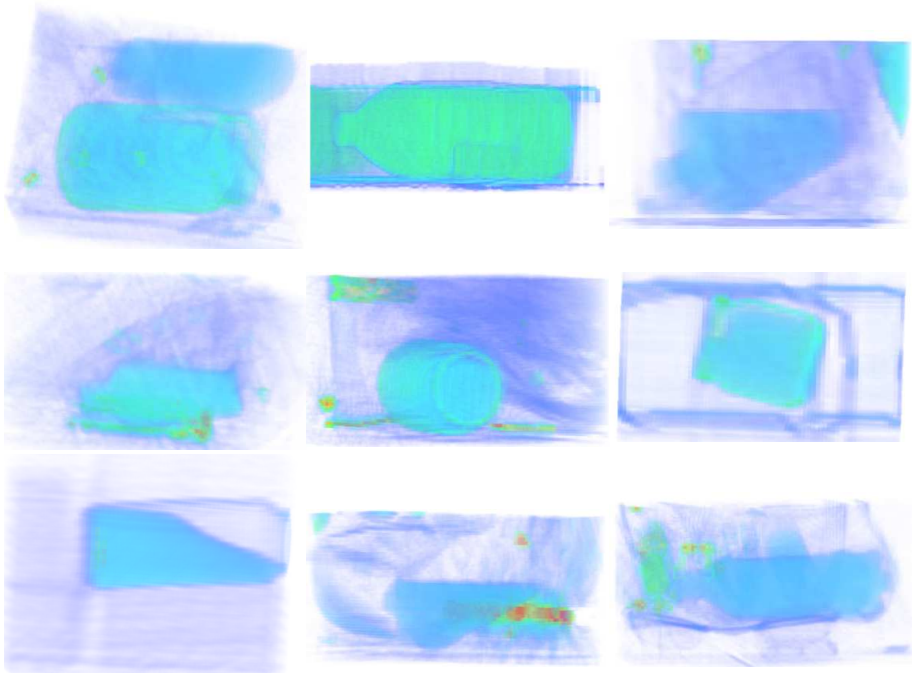
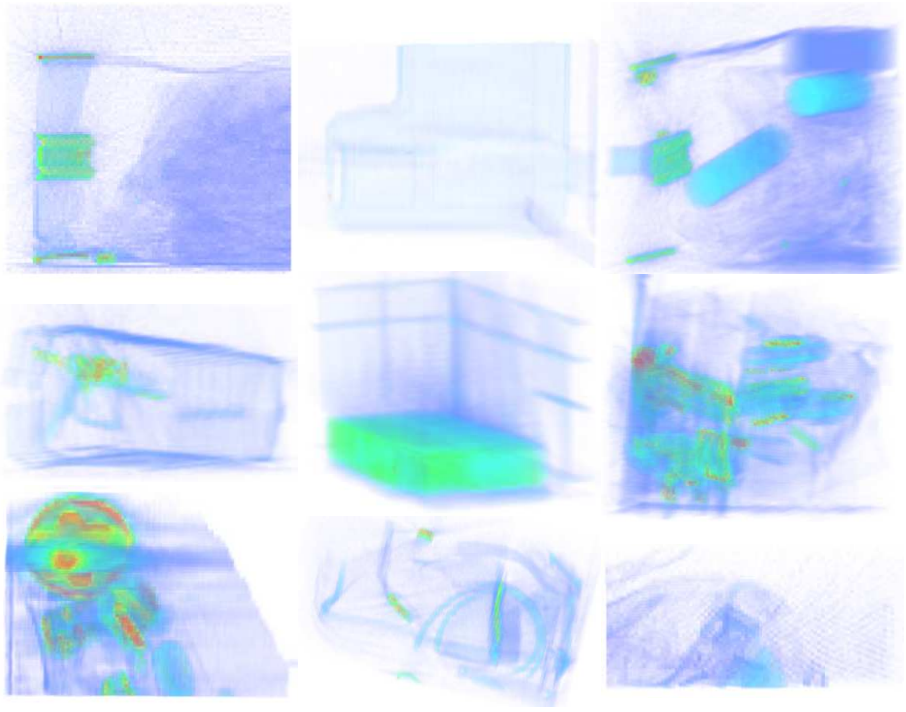


Figure 7.20: Incorrectly classified clutter as handguns



(a) Example bottles



(b) Example clear

Figure 7.21: Example volumes used for bottles experiment

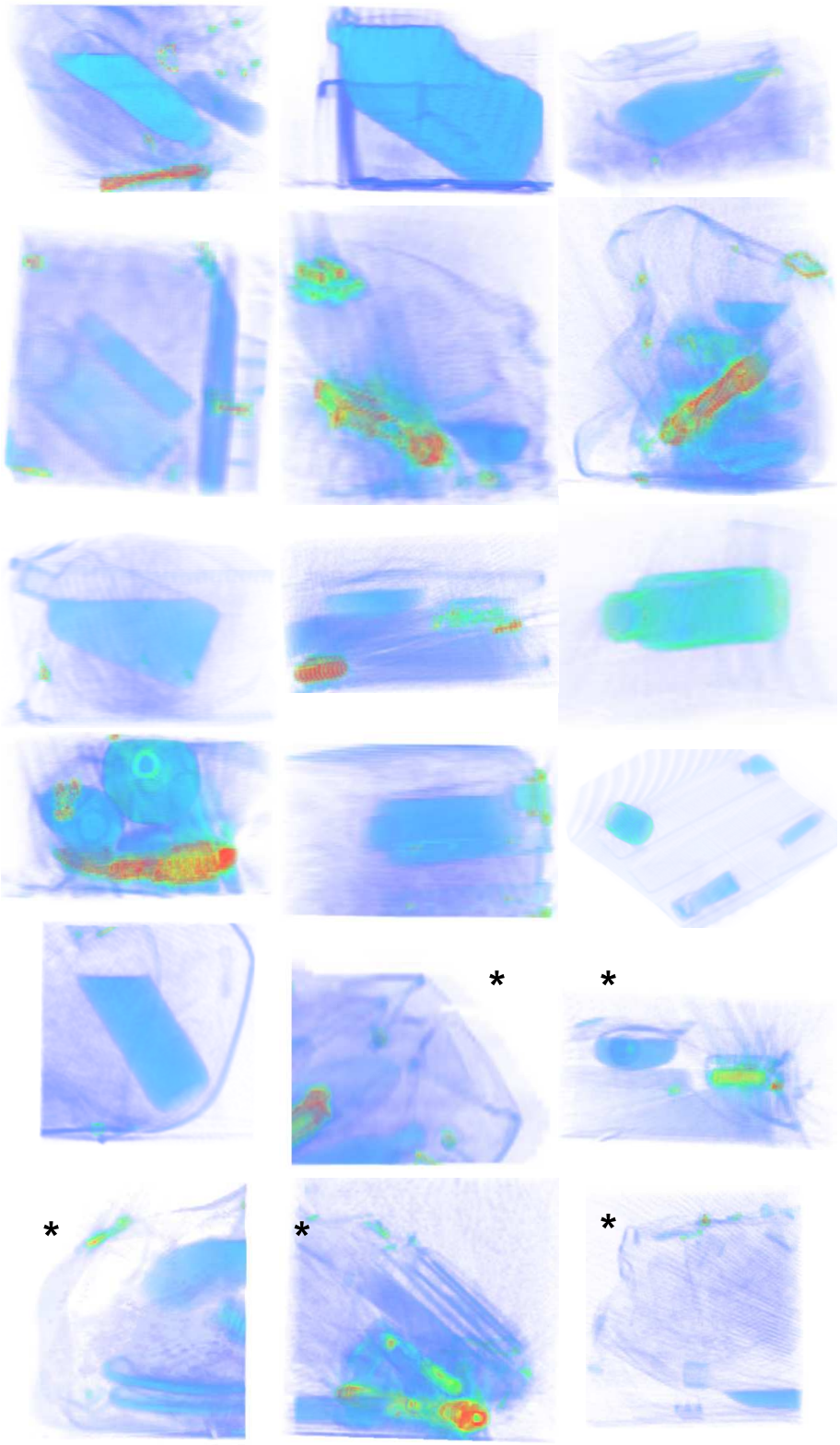


Figure 7.22: Incorrectly classified bottles (* partial bottles present)

Fold	True-positive rate (%)	False-positive rate (%)	Precision	Recall
0	98.1 (52/53)	0.0 (0/117)	1.000	0.981
1	98.1 (52/53)	0.0 (0/117)	1.000	0.981
2	92.5 (49/53)	0.9 (1/117)	0.980	0.925
3	96.2 (51/53)	0.0 (0/117)	1.000	0.962
4	98.1 (52/53)	1.7 (2/117)	0.963	0.981
5	90.6 (48/53)	1.7 (2/117)	0.960	0.906
6	100.0 (53/53)	0.0 (0/117)	1.000	1.000
7	98.1 (52/53)	0.0 (0/117)	1.000	0.981
8	94.3 (50/53)	5.1 (6/117)	0.893	0.943
9	100.0 (57/57)	0.9 (1/117)	0.983	1.000
Mean	96.6 ± 3.2 (517/534)	1.0 ± 1.6 (12/1170)	0.977 ± 0.018	0.966 ± 0.032

Table 7.4: Bottle sub-volume fold results

to the example liquids.

Overall we can again see from the examples shown in Figure 7.22, Figure 7.23 and Table 7.4 that the technique performs with very high precision and recall statistics matched by a mean 96.6% true-positive rate and mean 1.0% false-positive rate. This is performed over a realistic test set of baggage imagery as laid out in Section 6.4.1.

7.7 Conclusions

This work has shown that extending 2D visual-cortex models into 3D can achieve excellent results, exceeding those achieved using bag of features based on interest-point-descriptor methodologies (see Chapter 6). Table 7.5 shows the results achieved using the visual-cortex methodology when compared to the codebook approach of Chapter 6. Table 7.5a shows results achieved in handgun recognition where we see that the visual-cortex approach produces a similar recognition rate (96.8%) to the best codebook approach (97.3%) achieved using a density-histogram descriptor (see Section 6.5). The false-positive rate is lower for the cortex method (1.1%) compared to the density-histogram-codebook method (1.8%) though the measurement error is sufficient in both cases to prevent this being a significant difference. It is also noted that the precision result is higher for the cortex method (0.962) when compared to the best of the codebook approach (0.942, density histogram) indicating a slightly greater degree of confidence in the assertion by the cortex method of a true classification result.

Table 7.5b shows the results for recognition of bottles where we can see a distinct

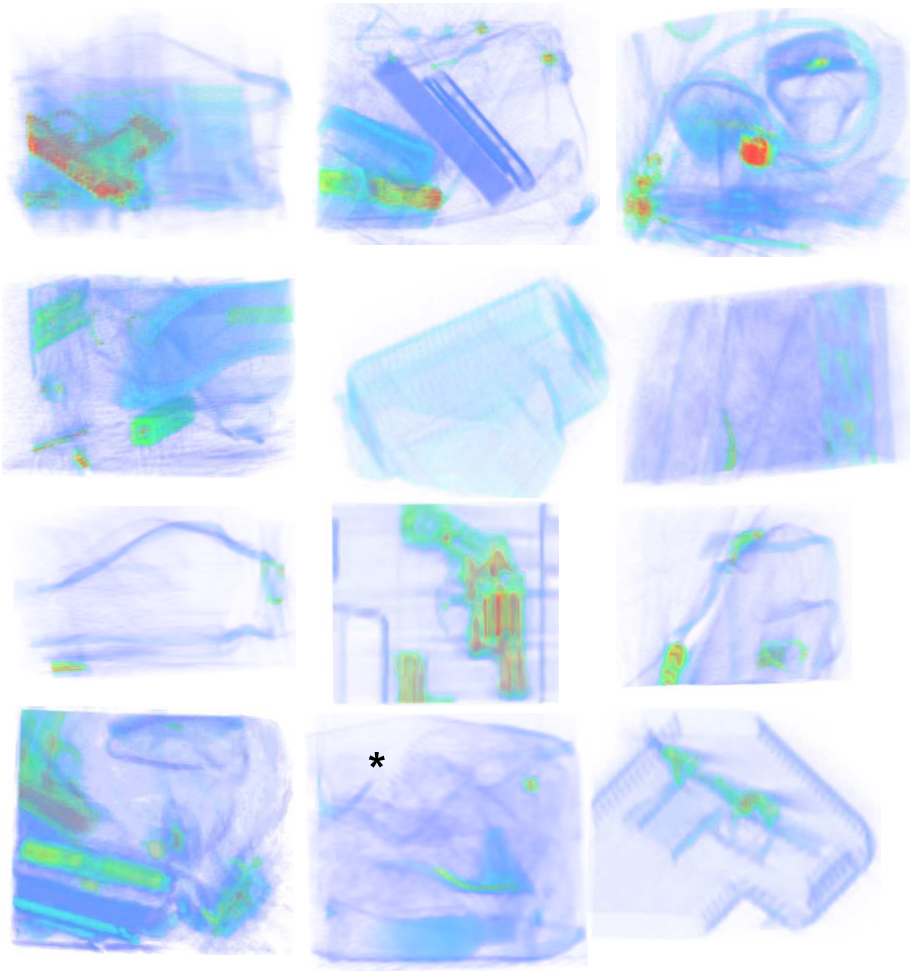


Figure 7.23: Clutter incorrectly classified as bottle (* partial bottle present)

Method	True-positive rate	False-positive rate	Precision	Recall
Visual Cortex	96.8 ± 2.6	1.1 ± 0.9	0.962±0.029	0.968±0.026
Codebook: SIFT	87.0 ± 5.4	3.8 ± 2.4	0.870±0.069	0.870±0.054
Codebook: RIFT	87.3 ± 3.9	5.1 ± 2.3	0.832±0.066	0.873±0.039
Codebook: DH	97.2 ± 3.4	1.8 ± 1.7	0.942±0.053	0.972±0.034
Codebook: DGH	97.2 ± 2.2	2.1 ± 1.2	0.932±0.035	0.972±0.022

(a) Handgun results

Method	True-positive rate	False-positive rate	Precision	Recall
Visual Cortex	96.6 ± 3.2	1.0 ± 1.6	0.977±0.034	0.966±0.032
Codebook: SIFT	82.8 ± 7.0	4.2 ± 1.2	0.900±0.025	0.828±0.070
Codebook: RIFT	78.3 ± 6.4	5.6 ± 2.4	0.864±0.052	0.783±0.064
Codebook: DH	89.3 ± 5.5	3.0 ± 1.4	0.932±0.029	0.893±0.055
Codebook: DGH	87.3 ± 6.8	4.0 ± 1.8	0.908±0.039	0.873±0.068

(b) Bottle results

Table 7.5: Comparison of visual cortex result with codebook results

difference between the recognition achieved using the visual-cortex methodology (96.6%) to the best codebook approach (89.3%, density histogram) even taking the measurement error into account. There is a more impressive difference in the false-positive rates with the visual-cortex methodology (1.0%) much lower than the best codebook method (3.0%, density histogram). We again see a higher value for the precision in the cortex result (0.977) when compared to the best of the codebook method (0.932, density histogram) indicating a greater confidence in each cortex-classification result.

It is interesting to note that the visual-cortex methodology results are similar for both handgun and bottle recognition tasks possible indicating that this approach is a more general solution to the task in hand. The codebook approach taken in Chapter 6 had a poorer performance in the recognition of bottles when compared to handguns indicating that tuning of that algorithm may be required depending on the target object type.

It is worth noting that the SIFT-descriptor results compares poorly to the visual-cortex approach for both true-positive rate, false-positive rate and precision.

Further work will include completion of the overall methodology of Mutch and Lowe (2008) in 3D. For example, Mutch and Lowe (2008) used patches of size $n \times n$ where $n = \{4,\ 8,\ 12,\ 16\}$. At present we have used $n = 4$. In the 2D case it was found

that best results were obtained using a random patch size selection in derivation of the classification patches (Section 7.4). When forming the S2 layer, modelling of likely neuron behaviour can be made by only retaining the dominant orientation results for each voxel in a patch rather than all orientations. This is coupled with an increase in the number of orientations modelled (4 becomes 12 for the 3D case) which could easily be extended in 3D. This has the added effect of reducing the amount of data stored in the S2 layer which would reduce the computation. Another step taken by Mutch and Lowe (2008) is to inhibit outputs from the S1/C1 layer as a result of a dominant neighbour orientation. This addition models 'lateral inhibition' in the visual cortex whereby the output level of one dominant neuron reduces or eliminates the output levels of its less dominant neighbours (Serre et al., 2005a).

Object localization within a large bag would be a task worthy of further work. At present we have investigated recognition of sub-volumes but we could imagine taking a complete baggage scan and locating an object through a sliding 3D volumetric window. Investigations using a greater number of object classes (mobile phones, knives for example) coupled with a larger amount of data is required to extend the work for the specific task of threat detection in baggage imagery.

Chapter 8

Conclusions

We now present a concluding summary of the achievements and results obtained during this research and propose areas that could be explored in the future.

8.1 Summary

- In Chapter 4 and Chapter 5 we demonstrated specific-instance recognition within complex 3D CT baggage imagery through the use of a 3D extension to the SIFT descriptor (Lowe, 2004) which extended prior 3D SIFT work that looked only at the problems of image registration/3D panorama creation (Allaire et al., 2008; Ni et al., 2009). We successfully demonstrated that recognition of complex metallic objects (revolver, pistol, binoculars) was possible but that reduced recognition rates seemed to follow as the object became smaller (iPod) or less dense (pistol frame). We also noted (Chapter 5) that the orientation of an object (pistol) as it enters the CT scanner greatly affects the resultant image and this can have a significant impact on the algorithm success. An investigation into keypoint matching examined the use of a fixed threshold, a fixed percentile or match distinction (Lowe, 2004) to determine the candidate match set. Best performance was achieved using the fixed-percentile approach. It was noted that part recognition using points of interest was unsuccessful using the Glock pistol frame as the SIFT keypoint locations were distorted/removed when the pistol barrel was attached.

- In Chapter 5 we compared the 3D extension to the SIFT descriptor from Chapter 4 against a number of alternative descriptor methodologies. A local density histogram and local density-gradient histogram proved to be adept as descriptors that did not require the generation of an invariant coordinate

system (a key part of the SIFT approach) and demonstrated near perfect performance in the recognition of a revolver and handgun in a complex 'whole-bag' situation. We also extended the RIFT descriptor (Lazebnik et al., 2005) into 3D and its performance was comparable to the SIFT descriptor with the advantage that its formulation was two orders of magnitude smaller (8 *vs.* 864 elements). This comparative work showed that improved recognition rates were possible using simpler invariant descriptors that do not require the generation of an invariant coordinate system (a key part of the SIFT approach).

- We then moved from the practical limitations of specific-instance recognition to the more applicable approach of object-class recognition. A rigorous investigation into the bag-of-features codebook methodology (Chapter 6) using a variety of descriptors, codebook-assignment methodologies and codebook size was performed. Two classes of object were considered (handguns, bottles) and we observed that the highest recognition rates (97.2% for handguns, 89.3% for bottles) were achieved using relatively simple descriptors (local density histogram; local density-gradient histogram) with the more complex 3D SIFT descriptor lagging in performance (87.0% for handguns, 82.8% for bottles) despite its superior reputation from application in 2D natural imagery. The results of these tests show a relative under-performance in the detection of sparse-feature objects (e.g. bottles) when compared to feature-rich (i.e. complex objects, e.g. handguns) indicating that the keypoint methodology employed is dependent on the generalized feature density of the objects considered.

- In Chapter 7 we examined an alternative approach to object-class recognition: a novel 3D extension of a visual-cortex-modelling methodology (Serre et al., 2005b; Mutch and Lowe, 2008). Successful detection of both handguns and bottles is high, indicating that this approach may be a more generalized solution to object class recognition in CT baggage imagery given both its detection performance and limited training data requirements. A direct comparison between this approach and that of the bag-of-features codebook methodology was performed using the same dataset for both. The results of this work showed that extending 2D visual cortex models into 3D can achieve excellent results (96.8% for handguns, 96.6% for bottles), comparable to those achieved using interest point based bag-of-features methodologies (see Chapter 6) for handgun recognition and significantly better for bottle detection. Of particular note is the lower false-positive rates that accompany the visual-cortex

method ($\approx 1.0\%$) compared to the interest point approaches ($2.0\% - 4.0\%$). These results are a compelling finding that leads us to believe that ultimately this approach could result in a more generic recognition solution for complex 3D imagery.

A word of caution is required for all these results and observations. It was not possible to obtain a large dataset for this work; free and easy access to the CT scanner was not feasible. This resulted in a relatively small amount of captured data ($\approx$500 baggage items). Consequently we need to be circumspect in regards of the obtained results. As an example, it was only possible to use two classes of object for the work in Chapter 6 and Chapter 7. Standard work in 2D imagery uses many more classes and is freely available for testing around the world. For instance, the Caltech-101 database has 101 classes of object for recognition algorithms to be tested upon (Fei-Fei et al., 2007). No such database exists for 3D CT baggage data. The class recognition results in Chapter 6 and Chapter 7 include standard deviation measurements and these reinforce this point: in many cases it cannot be asserted that one technique outperforms another simply because, when including the measurement standard deviation, a clear distinction is not present.

These results present us with a number of areas for future research that range in scope and complexity. We will now propose areas that could be explored in the future.

8.2 Future work

In all cases future work would benefit from a larger dataset in terms of both known target items, classes of item and clutter baggage. The data captured for this work was generated using a number of baggage items (rucksacks, suitcases etc) that had been packed by a small number of people. It is likely that these may not be representative of genuine baggage items; very little food was packed, for example. The resolution of this issue would involve the capture of actual baggage items at an airport, though the logistics in organizing this would be considerable.

We chose to explore the recognition problem using isotropic voxels of size $2.5mm^3$ to ease the algorithm-development path. We could extend all aspects of the work covered in this thesis to function with the original anisotropic voxels of size $1.6mm \times 1.6mm \times 5mm$ and determine subsequent performance as this would speed execution of the algorithm. The 3D-SIFT work of Allaire et al. (2008) was performed using anisotropic voxels with reasonable results. We would anticipate that a move to anisotropic voxels would not significantly alter our results. For the work explored

in this thesis we operated the CT scanner in a static-scanning mode rather than its default helical-scan mode. Exploring performance on imagery from helical scans would be of interest as the scanning process is faster in this mode of operation.

The overall image quality is poor; metallic artefacts and noise degrade the fidelity. An investigation into computer vision based enhancement techniques is currently ongoing and a review of recognition performance from any of the approaches in this thesis on any improved imagery would be of interest.

At present all research has been performed on CT volumes for which each voxel is related to the local density at that point in the baggage item. An alternative imagery paradigm is the formation of material-type volumes that can be derived from the dual-energy CT scan data (high and low energy slice images, see Chapter 3). The application of the computer vision techniques to this alternative form of imagery would add a novel aspect to recognition within this imaging domain. The combination of material type and density volumetric data may enhance recognition performance and would be worthy of further investigation.

We will now detail specific areas of future work from each recognition approach taken in this thesis.

8.2.1 Specific-instance approach

Specific-instance recognition has limited practical application in baggage understanding. There are a few areas of work in this area, though the class-based approaches are of more interest:

- The underperformance of the SIFT descriptor (Chapter 4, Chapter 5) is believed to be due to disruption of the methodology that derives an orientation-invariant descriptor; inconsistent formulation of the invariant coordinate set will hinder the feature-matching process. This disruption is likely down to the presence of metallic artefacts and noise within the imagery. One area of future work could be the derivation of an improved dominant-orientation methodology for the 3D SIFT descriptor that copes better within the CT imagery.

- In the comparison of descriptors the Difference of Gaussian method was chosen to determine the points of interest. This method may not be the best approach for the CT-baggage imagery and an investigation into alternative keypoint-selection methodologies could be performed.

- We investigated a number of descriptor methodologies but the use of a 3D extension to SURF (Bay et al., 2008; Knopp et al., 2010) was not attempted.

This is an obvious area for further research as the SURF descriptor also relies of a dominant direction method in its derivation that may be disrupted by the CT artefacts and limit its applicability to this type of imagery.

8.2.2 Codebook approach

This work could be extended as follows:

- The bag-of-features methodology does not consider object geometry in its recognition approach. Extending this approach to include the relative position of features is likely to enhance recognition.

Other, less important areas for work include:

- At present we have demonstrated recognition of extracted sub-volumes. An attempt was made to perform whole-volume recognition (see Appendix B) which demonstrated reasonable true-positive detection but noticeably high false-positive rates. The addition of techniques that provide object localization within the codebook approach would aid human operators. This could be achieved by examining sub-volumes using a moving volumetric window to determine if the false-positive rates can be reduced for the examination of a whole bag.

- Our work into codebook-based class recognition examined a number of assignment methodologies. The uncertainty-assignment approach (Section 6.3.3), though producing consistently good results, is time consuming ($\approx O(K^2)$ for K clusters). An investigation into the performance of alternatives which require less computation would be of interest. Simpler approaches that achieve similar performance could be investigated, for example only assigning to the closest N clusters rather than all clusters (Philbin et al., 2008).

- The results quoted within Chapter 6 minimized the classification errors during training in order to configure the SVM classifier. The result of this is a single-classification result (true and false-positive rates) but an examination of performance using ROC plots (as in Chapter 5) for the results would be of interest as it would show how the true-positive-rate performance varies with the false-positive rate allowing the tuning of the configuration for operation in the field.

8.2.3 Visual-cortex approach

The visual-cortex methodology of Chapter 7 has demonstrated great promise in the class-recognition task. There are a number of aspects of this approach that deserve further investigation.

Major aspects include:

- At present little work has been performed on optimizing parameter settings. For example, the number of 3D Gabor-filter directions (currently 10), the number of classification features (1500) and the number of scale-space levels (10) could all be varied to determine the most favourable values.

- The choice of classification feature size was limited in our research to a $4 \times 4 \times 4$ voxel size. In Mutch and Lowe (2008) best results were obtained by starting the selection process with randomly selected patches of size $n \times n$, $n \in \{4, 8, 12, 16\}$. An implementation of this within the 3D extension could further enhance recognition performance.

- The work of Mutch and Lowe (2008) was completed with the introduction of a number of biologically plausible additions to the visual-cortex hierarchy. One example is the introduction of sparsity to the S2 layer by only retaining the dominant Gabor response rather than all 10 responses. Another example is the simulation of lateral inhibition on the outputs of the S1 and C1 layers which can force a S1/C1 output to zero.

- Work on object localisation using a moving sub-volume window would explore the performance in detection within a whole-bag scenario.

Other areas include:

- The feature-selection step was shown to improve recognition performance (Mutch and Lowe, 2008). There is the opportunity to investigate whether the methodology employed (selection by SVM normal component magnitude) is the best or whether alternatives would yield improved performance.

- An examination of performance using ROC plots (as in Chapter 5) for the results would also be of interest for this approach.

- At present the implementation is not optimized for speed in either its algorithm or use of hardware. An investigation into the use of GPUs to enhance performance would be required to achieve an appropriate speed of detection for deployment in an airport scenario.

- The application of this technique to other imagery modalities is an engaging proposition. Its application to medical imagery (CT and MRI) is one possibility as is its use in action/gesture recognition within spatio-temporal volumetric data. It may be easier to compare the performance of the visual-cortex approach against other techniques within an imaging domain for which data is more readily available.

References

S. Allaire, J. Kim, S. Breen, D. Jaffray, and V. Pekar. Full orientation invariance and improved feature selectivity of 3D SIFT with application to medical image analysis. In *IEEE computer society conference on computer vision and pattern recognition workshops*, pages 1–8, 2008.

J. M. Aman, J. Yao, and R. M. Summers. Content-based image retrieval on CT colonography using rotation and scale invariant features and bag-of-words model. In *7th IEEE international symposium on biomedical imaging: from nano to macro*, pages 1357–1360, 2010.

M. Antonelli, M. Cococcioni, B. Lazzerini, and F. Marcelloni. Computer-aided detection of lung nodules based on decision fusion techniques. *Pattern analysis and applications*, 14(3):295–310, 2011.

D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms*, pages 1027–1035. Society for industrial and applied mathematics, 2007.

K. Arun, T. Huang, and S. Blostein. Least-squares fitting of two 3-D point sets. *IEEE transactions on pattern analysis and machine intelligence*, 9(5):698–700, 1987.

M. Baştan, M. Yousefi, and T. Breuel. Visual words on baggage x-ray images. In P. Real, D. Diaz-Pernil, H. Molina-Abril, A. Berciano, and W. Kropatsch, editors, *Computer analysis of images and patterns*, volume 6854 of *Lecture notes in computer science*, pages 360–368. Springer Berlin / Heidelberg, 2011.

D. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern recognition*, 13(2):111–122, 1981.

D. H. Ballard and C. M. Brown. *Computer vision*. Prentice-Hall Inc., 1982.

J. Barrett and N. Keat. Artifacts in CT: recognition and avoidance. *Radiographics*, 24(6):1679–1691, 2004.

H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer vision and image understanding*, 110(3):346–359, 2008.

C. Belcher and Y. Du. Region-based SIFT approach to iris recognition. *Optics and lasers in engineering*, 47(1):139–147, 2009.

W. Bi, Z. Chen, L. Zhang, and Y. Xing. A Volumetric Object Detection Framework with Dual-Energy CT. In *IEEE Nuclear Science Symposium Conference Record*, pages 1289–1291, 2008.

W. Bi, Z. Chen, L. Zhang, and Y. Xing. Fast detection of 3d planes by a single slice detector helical ct. In *IEEE nuclear science symposium conference record*, pages 954–955, 2009.

M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli. On the use of sift features for face authentication. In *IEEE conference on computer vision and pattern recognition workshop*, page 35, 2006.

C. Bishop. *Pattern recognition and machine learning*. Springer New York, 2006.

A. Bosch, A. Zisserman, and X. Muoz. Image classification using random forests and ferns. In *IEEE 11th international conference on computer vision*, pages 1–8, 2007.

M. Brown and D. Lowe. Invariant features from interest point groups. In *British machine vision conference*, pages 656–665, 2002.

M. Brown and D. Lowe. Automatic panoramic image stitching using invariant features. *International journal of computer vision*, 74(1):59–73, 2007.

J. Bustard and M. Nixon. Robust 2d ear registration and recognition based on sift point matching. In *2nd IEEE international conference on biometrics: theory, applications and systems*, pages 1–6, 2008.

J. Chan, A. Omar, J. Evans, D. Downes, X. Wang, and Y. Liu. Feasibility of sift to synthesise kdex imagery for aviation luggage security screening. In *3rd international conference on crime detection and prevention*, pages 1–6, 2009.

W. Cheung and G. Hamarneh. N-Sift: N-dimensional scale invariant feature transform for matching medical images. *4th IEEE international symposium on biomedical imaging: from nano to macro*, pages 720–723, 2007.

O. Chum and J. Matas. Matching with PROSAC - progressive sample consensus. In *IEEE computer society conference on computer vision and pattern recognition*, pages 220–226, 2005.

O. Chum, J. Matas, and J. Kittler. Locally optimized ransac. In *DAGM symposium*, pages 236–243, 2003.

A. Collet, D. Berenson, S. S. Srinivasa, and D. Ferguson. Object recognition and full pose registration from a single image for robotic manipulation. In *IEEE international conference on robotics and automation*, pages 48–55, 2009.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *European conference on computer vision workshop on statistical learning in computer vision*, pages 1–22, 2004.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, volume 1, pages 886–893, 2005.

R. Dalvi, I. Hacihaliloglu, and R. Abugharbieh. 3D ultrasound volume stitching using phase symmetry and Harris corner detection for orthopaedic applications. In B. M. Dawant and D. R. Haynor, editors, *Medical Imaging 2010: Image Processing*, volume 7623, page 762330. SPIE, 2010.

R. Desimone, T. Albright, C. Gross, and C. Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. *The journal of neuroscience*, 4(8): 2051–2062, 1984.

R. Desimone, S. J. Schein, J. Moran, and L. G. Ungerleider. Contour, color and shape analysis beyond the striate cortex. *Vision research*, 25(3):441–452, 1985.

K. Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4): 198–211, 2007.

M. Donoser and H. Bischof. 3d segmentation by maximally stable volumes (msvs). In *18th international conference on pattern recognition.*, volume 1, pages 63–66, 2006.

C. Evans. Notes on the openSURF library. Technical report CSTR-09-001. Technical report, Bristol University, 2009.

J. Evans. Kinetic depth effect x-ray (kdex) imaging for security screening. In *International conference on visual information engineering*, pages 69–72, 2003.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (VOC) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

T. Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8): 861–874, 2006.

L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. *Computer vision and image understanding*, 106(1):59–70, 2007.

P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International journal of computer vision*, 61(1):55–79, 2005.

P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE conference on computer vision and pattern recognition*, pages 1–8, 2008.

P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

Q. Feng, M. Foskey, S. Tang, W. Chen, and D. Shen. Segmenting ct prostate images using population and patient-specific statistics for radiotherapy. In *IEEE international symposium on biomedical imaging: from nano to macro*, pages 282–285, 2009.

M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE transactions on computers*, C-22(1):67–92, 1973.

R. B. Fisher. *From surfaces to objects: computer vision and three dimensional scene analysis*. John Wiley and Sons, 1989.

R. B. Fisher. Performance comparison of ten variations on the interpretation-tree matching algorithm. In *European conference on computer vision*, volume 1, pages 507–512, 1994.

G. Flitton, T. Breckon, and N. Megherbi. Object recognition using 3D SIFT in complex CT volumes. In F. Labrosse, R. Zwiggelaar, Y. Liu, and B. Tiddeman, editors, *Proceedings of the British machine vision conference*, pages 11.1–11.12. BMVA Press, 2010.

W. Forstner. A feature based correspondence algorithm for image matching. *International archives of photogrammetry and remote sensing*, 26(3):150–166, 1986.

R. Gesick, C. Saritac, and C.-C. Hung. Automatic image analysis process for the detection of concealed weapons. In *Proceedings of the 5th annual workshop on cyber security and information intelligence research*, pages 1–4, 2009.

W. E. L. Grimson and T. Lozano-Perez. Localizing overlapping parts by searching the interpretation tree. *IEEE transactions on pattern analysis and machine intelligence*, 9(4):469–482, 1987.

M. Grundmann, F. Meier, and I. Essa. 3d shape context and distance transform for action recognition. In *19th international conference on pattern recognition*, pages 1–4, 2008.

C. Harris and M. Stephens. A combined corner and edge detector. In *Fourth Alvey vision conference*, pages 147–151, 1988.

K. Huang, D. Tao, Y. Yuan, X. Li, and T. Tan. Biologically inspired features for scene classification in video surveillance. *IEEE transactions on systems, man, and cybernetics, part B: cybernetics*, 41(1):307–313, 2011.

D. Hubel and T. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The journal of physiology*, 148(3):574–591, 1959.

D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The journal of physiology*, 160(1):106–154, 1962.

D. Hubel and T. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The journal of physiology*, 195(1):215–243, 1968.

H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *IEEE 11th international conference on computer vision*, pages 1–8, 2007.

F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Tenth IEEE international conference on computer vision*, volume 1, pages 604–610, 2005.

A. C. Kak and M. Slaney. *Principles of computerized tomographic imaging*. IEEE Press, 1988.

J. Knopp, M. Prasad, G. Willems, R. Timofte, and L. Van Gool. Hough transform and 3D SURF for robust three dimensional classification. In *11th European conference on computer vision*, volume 6, pages 589–602, 2010.

J. J. Koenderink and A. J. van Doorn. Surface shape and curvature scales. *Image and vision computing*, 10(8):557–564, 1992.

Y. Lamdan, J. Schwartz, and H. Wolfson. Object recognition by affine invariant matching. In *Computer society conference on computer vision and pattern recognition*, pages 335–344, 1988.

I. Laptev. On space-time interest points. *International journal of computer vision*, 64(2):107–123, 2005.

S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE transactions on pattern analysis and machine intelligence*, 27(8):1265–1278, 2005.

S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *IEEE computer society conference on computer vision and pattern recognition,*, volume 2, pages 2169–2178, 2006.

B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International journal of computer vision*, 77(1): 259–289, 2008.

D. Lewis. Naive (bayes) at forty: the independence assumption in information retrieval. In *Proceedings of the 10th European Conference on Machine Learning*, pages 4–15. Springer, 1998.

Y. Liu, J. Yang, D. Zhao, and J. Liu. Computer aided detection of lung nodules based on voxel analysis utilizing support vector machines. In *International conference on future biomedical information engineering*, pages 90–93, 2009.

D. Lowe. The viewpoint consistency constraint. *International journal of computer vision*, 1(1):57–72, 1987.

D. Lowe. Object recognition from local scale-invariant features. In *The proceedings of the seventh IEEE international conference on computer vision.*, volume 2, pages 1150–1157, 1999.

D. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

J. Luo, Y. Ma, E. Takikawa, S. Lao, M. Kawade, and B. Lu. Person-specific sift features for face recognition. In *IEEE international conference on acoustics, speech and signal processing*, volume 2, pages 593–596, 2007.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. University of California Press, 1967.

P. Mahalanobis. On the generalized distance in statistics. In *Proceedings of the national institute of science, Calcutta*, volume 12, pages 49–55, 1936.

S. Maji, A. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *IEEE conference on computer vision and pattern recognition*, pages 1–8, 2008.

J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767, 2004.

P. McIlroy, E. Rosten, S. Taylor, and T. Drummond. Deterministic sample consensus with multiple match hypotheses. In *Proceedings of the British machine vision conference*, pages 111.1–111.11. BMVA Press, 2010.

N. Megherbi, G. Flitton, and T. Breckon. A classifier based approach for the detection of potential threats in CT based baggage screening. In *Proceedings of the IEEE international conference on image processing*, pages 1833–1836, 2010.

K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 27(10):1615–1630, 2005.

K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Gool. A comparison of affine region detectors. *International journal of computer vision*, 65(1):43–72, 2005.

D. Mladenić, J. Brank, M. Grobelnik, and N. Milic-Frayling. Feature selection using linear classifier weights: interaction with classification models. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval*, pages 234–241. ACM, 2004.

J. Mutch and D. Lowe. Multiclass object recognition with sparse, localized features. In *IEEE computer society conference on computer vision and pattern recognition*, volume 1, pages 11–18, 2006.

J. Mutch and D. Lowe. Object class recognition and localization using sparse features with limited receptive fields. *International journal of computer vision*, 80(1):45–57, 2008.

S. Nercessian, K. Panetta, and S. Agaian. Automatic detection of potential threat objects in x-ray luggage scan images. In *IEEE conference on technologies for homeland security*, pages 504–509, 2008.

D. Ni, Y. Chui, Y. Qu, X. Yang, J. Qin, T. Wong, S. Ho, and P. Heng. Reconstruction of volumetric ultrasound panorama based on improved 3D SIFT. *Computerized medical imaging and graphics*, 33(7):559–566, 2009.

M. Niemeijer, M. K. Garvin, K. Lee, B. V. Ginneken, M. D. Abràmoff, and M. Sonka. Registration of 3D spectral OCT volumes using 3D SIFT feature point matching. In *Medical imaging 2009: image processing*, volume 7259, page 72591I. SPIE, 2009.

D. Nistér. Preemptive ransac for live structure and motion estimation. *Machine vision and applications*, 16(5):321–329, 2005.

M. Novotni and R. Klein. Shape retrieval using 3d zernike descriptors. *Computer-aided design*, 36(11):1047–1062, 2004.

E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *9th European conference on computer vision*, volume 3954, pages 490–503. Springer, 2006.

C. Oertel and P. Bock. Identification of objects-of-interest in x-ray images. In *35th IEEE applied imagery and pattern recognition workshop*, page 17, 2006.

J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: improving particular object retrieval in large scale image databases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–8, 2008.

H. Riemenschneider, M. Donoser, and H. Bischof. Bag of optical flow volumes for image sequence recognition. In *Proceedings of the British machine vision conference*, 2009.

M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2:1019–1025, 1999.

M. Riesenhuber and T. Poggio. How visual cortex recognizes objects: the tale of the standard model. In *The visual neurosciences*, volume 2, pages 1640–1653. MIT Press, 2003.

E. Rosten and T. Drummond. Machine learning for high-speed corner detection. *Lecture notes in computer science*, 3951:430, 2006.

Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

T. Sattler, B. Leibe, and L. Kobbelt. Scramsac: improving ransac's efficiency with a spatial consistency filter. In *IEEE 12th international conference on computer vision*, pages 2090–2097, 2009.

C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 19(5):530–535, 1997.

M. Schuckers. Receiver operating characteristic and equal error rate. In *Computational methods in biometric authentication*, Information science and statistics, chapter 5, pages 155–204. Springer London, 2010.

W. Schweizer. *Numerical quantum dynamics*, volume 9. Kluwer academic publishers, 2001.

P. Scovanner, S. Ali, and M. Shah. A 3-dimensional SIFT descriptor and its application to action recognition. In *Proceedings of the 15th international conference on multimedia*, pages 357–360. ACM Press New York, 2007.

S. Se, D. Lowe, and J. Little. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *The international journal of robotics research*, 21(8):735–758, 2002.

T. Serre, L. Wolf, and T. Poggio. A new biologically motivated framework for robust object recognition. (cbcl paper 239 / ai memo 2004-017). Technical report, Massachusetts institute of technology, Cambridge, MA., 2004.

T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. Technical report, Massachusetts institute of technology, Center for biological and computational learning, 2005a.

T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *IEEE computer society conference on computer vision and pattern recognition*, volume 2, pages 994–1000, 2005b.

T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE transactions on pattern analysis and machine intelligence*, pages 411–426, 2007.

N. E. L. Shanks and A. L. W. Bradley. *Handbook of checked baggage screening: advanced airport security operation*. Wiley, 2nd edition, 2004.

J. Shi and C. Tomasi. Good features to track. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, pages 593–600, 1994.

R. Sim, P. Elinas, M. Griffin, and J. Little. Vision-based slam using the rao-blackwellised particle filter. In *IJCAI workshop on reasoning with uncertainty in robotics*, pages 9–16, 2005.

J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of the international conference on computer vision*, volume 2, pages 1470–1477, 2003.

J. Sivic, M. Everingham, and A. Zisserman. Person spotting: video shot retrieval for face sets. In *4th international conference on image and video retrieval*, pages 226–236. Springer, 2005.

I. Sluimer, A. Schilham, M. Prokop, and B. van Ginneken. Computer analysis of computed tomography scans of the lung: a survey. *IEEE transactions on medical imaging*, 25(4):385–405, 2006.

K. Suzuki, M. Epstein, I. Sheu, R. Kohlbrenner, D. Rockey, and A. Dachman. Massive-training artificial neural networks for cad for detection of polyps in ct colonography: false-negative cases in a large multicenter clinical trial. In *5th IEEE international symposium on biomedical imaging: from nano to macro*, pages 684–687, 2008.

R. Szeliski. *Computer vision: algorithms and applications*. Springer-Verlag New York Inc, 2010.

H. Tamimi, H. Andreasson, A. Treptow, T. Duckett, and A. Zell. Localization of mobile robots with omnidirectional vision using particle filter and iterative sift. *Robotics and autonomous systems*, 54(9):758–765, 2006.

T. Tommasini, A. Fusiello, E. Trucco, and V. Roberto. Making good features track better. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition*, pages 178–183, 1998.

P. Torr and A. Zisserman. Mlesac: a new robust estimator with application to estimating image geometry. *Computer vision and image understanding*, 78(1): 138–156, 2000.

Z. Tu. Probabilistic boosting-tree: learning discriminative models for classification, recognition, and clustering. In *IEEE international conference on computer vision*, volume 2, pages 1589–1596, 2005.

Z. Tu, X. Zhou, A. Barbu, L. Bogoni, and D. Comaniciu. Probabilistic 3d polyp detection in ct images: the role of sample alignment. In *IEEE computer society conference on computer vision and pattern recognition*, volume 2, pages 1544–1551, 2006.

M. Urschler, J. Bauer, H. Ditt, and H. Bischof. Sift and shape context for feature-based nonlinear registration of thoracic ct images. In *Computer vision approaches to medical image analysis*, volume 4241 of *Lecture notes in computer science*, pages 73–84. Springer Berlin/Heidelberg, 2006.

J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *IEEE transactions on pattern analysis and machine intelligence*, 32(7):1271–1283, 2010.

L. Vincent and P. Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE transactions on pattern analysis and machine intelligence*, pages 583–598, 1991.

P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, volume 1, pages 511–518, 2001.

D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional selection for object recognition-a gentle way. In *Second International Workshop on Biologically Motivated Computer Vision*, pages 472–479. Springer, 2002.

G. Wang and M. Vannier. Stair-step artifacts in three-dimensional helical CT: an experimental study. *Radiology*, 191(1):79–83, 1994.

G. Wang, D. Snyder, and M. Vannier. Iterative deblurring for CT metal artifact reduction. *IEEE transactions on medical imaging*, 15(5):657–664, 1996.

J. Yang, Y. Jiang, A. Hauptmann, and C. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on multimedia information retrieval*, pages 197–206, 2007.

Z. Yi, C. Zhiguo, and X. Yang. Multi-spectral remote image registration based on sift. *Electronics letters*, 44(2):107–108, 2008.

Z. Ying, R. Naidu, K. Guilbert, D. Schafer, and C. Crawford. Dual energy volumetric x-ray tomographic sensor for luggage screening. In *IEEE sensors applications symposium*, pages 1–6, Feb. 2007.

H. Yoshida, J. Näppi, P. Maceneaney, D. Rubin, and A. Dachman. Computer-aided diagnosis scheme for detection of polyps at ct colonography. *Radiographics*, 22(4): 963–979, 2002.

T.-H. Yu, O. Woodford, and R. Cipolla. An evaluation of volumetric interest points. In *International conference on 3D imaging, modeling, processing, visualization and transmission*, pages 282–289, 2011.

S. Zhao, D. Robeltson, G. Wang, B. Whiting, and K. Bae. X-ray CT metal artifact reduction using wavelets: an application for imaging total hip prostheses. *IEEE transactions on medical imaging*, 19(12):1238–1247, 2000.

Appendix A

Codebook: bottle sub-volume results

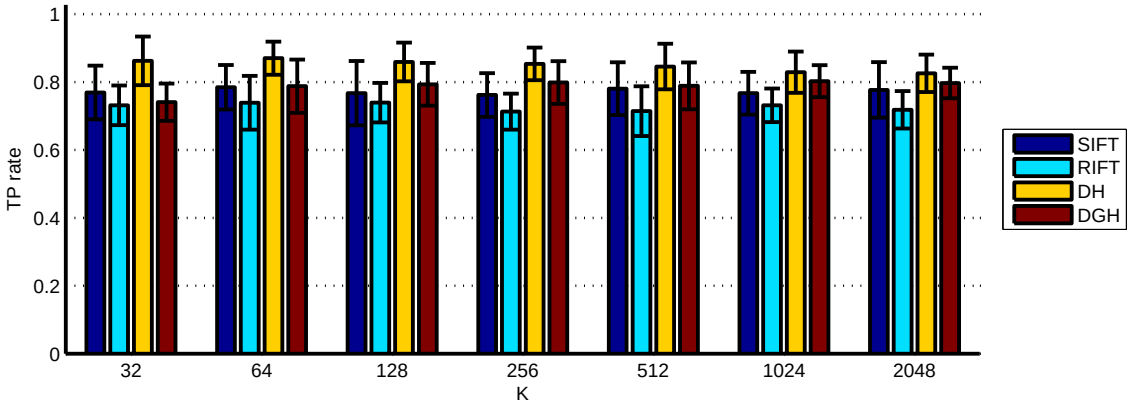
Here we present a detailed analysis of the results obtained using bottle sub-volumes as the target item in the codebook approach discussed in Chapter 6.

A.1 Hard assignment

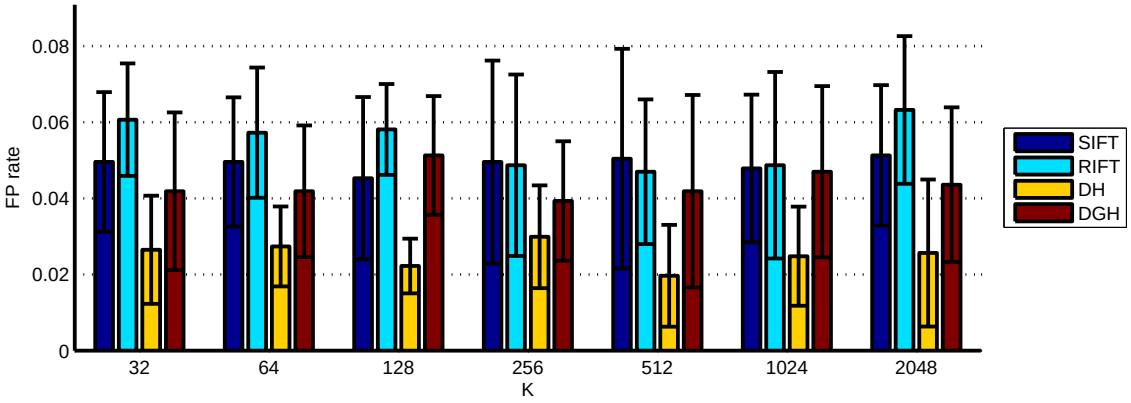
We again examine the hard-assignment methodology whilst varying the number of clusters is given by $K = 2^n$ for $n = \{6, 7, 8, 9, 10, 11\}$. Figure A.1 shows the true-positive and false-positive results for each descriptor type as the number of clusters is varied. The measured results show little variation as the number of clusters is increased but there is a distinct difference in performance when we consider the descriptors. The least effective result is given by RIFT followed by SIFT and DGH with the DH descriptor achieving the highest true-positive rate for all settings of K. The largest true-positive result is obtained using the DH descriptor with a value of 87.0% with a low false-positive rate of 2.7%. Table A.1 shows the settings for each descriptor that achieved the highest true-positive results. In all cases the results are significantly lower than for handguns (Table 6.2). For example, detection of handguns is 96.1% compared to 87.0% for bottles when using the density-histogram descriptor. For the density-gradient-histogram descriptor the difference is greater: 97.2% for handguns; 80.3% for bottles. Contrasting the true-positive detection rates, the false-positive rates are also higher than handgun detection again reflecting a poorer recognition system.

A.2 Kernel assignment

Kernel-assignment performance using the SIFT descriptor is shown in Figure A.2. The chart of true-positive rates in Figure A.2a is similar to that obtained for hand-



(a) True-positive performance

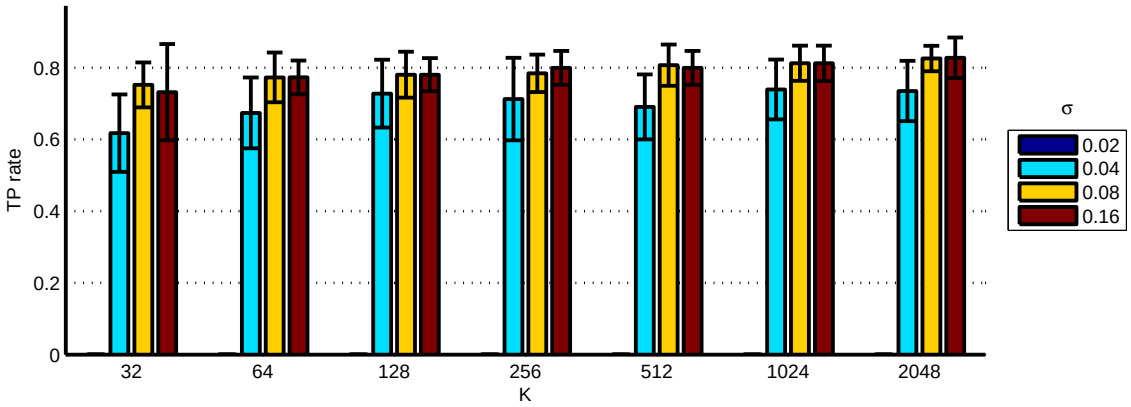


(b) False-positive performance

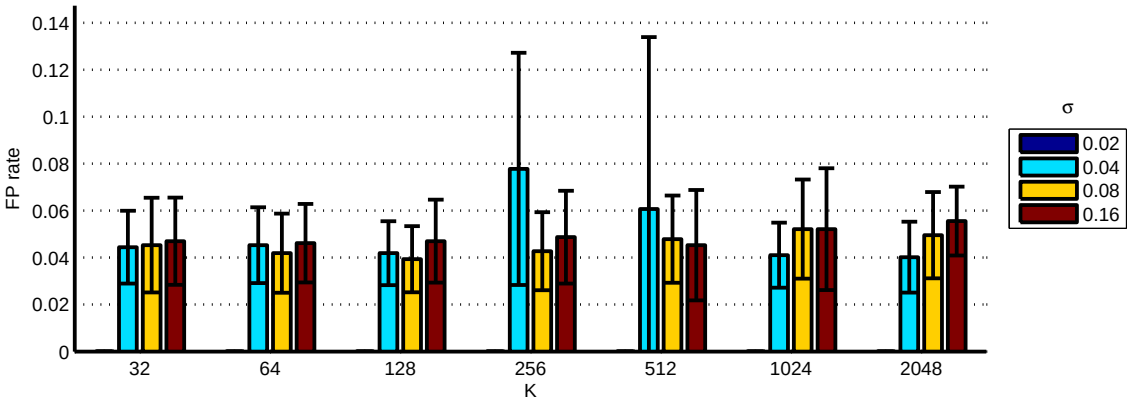
Figure A.1: Handgun sub-volume results using hard-assignment

Descriptor	K	TP rate (%)	FP rate (%)
SIFT	64	78.5 ± 6.6	5.0 ± 1.7
RIFT	128	73.9 ± 5.8	5.8 ± 1.2
DH	64	87.0 ± 4.9	2.7 ± 1.1
DGH	1024	80.3 ± 4.7	4.7 ± 2.3

Table A.1: Handgun sub-volume best detection rates for each descriptor using hard assignment



(a) True-positive performance

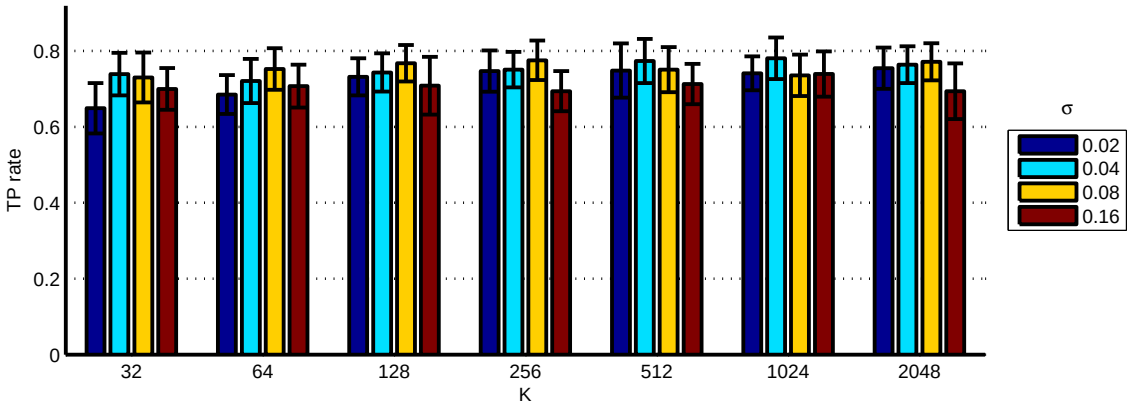


(b) False-positive performance

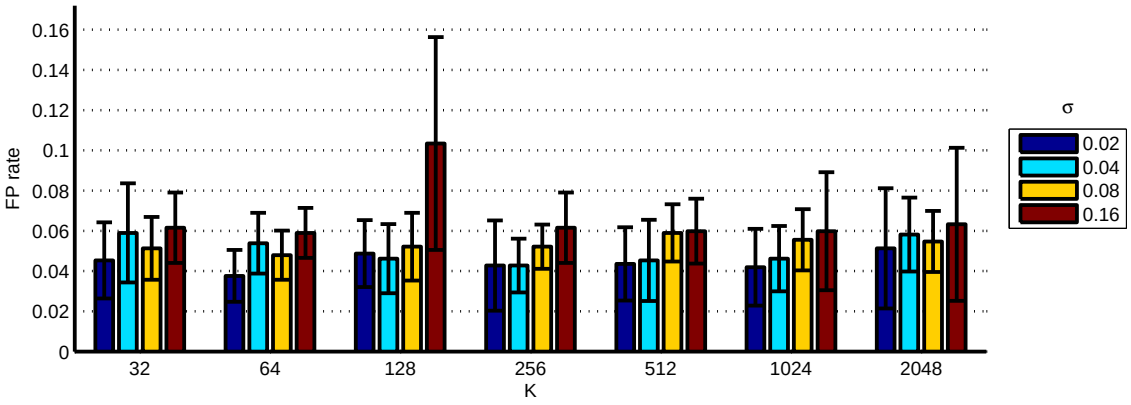
Figure A.2: Bottle sub-volume results using kernel assignment, SVM classification for SIFT descriptor

guns (Figure 6.12a) where we can see that best performance for higher values of both the smoothing parameter, σ, and number of codewords, K. In this case the highest detection rate is 82.8% for ($K = 2048$, $\sigma = 0.16$) with a corresponding false-positive rate of 5.6%. These results are similar to those achieved in handgun recognition (TPR: 85.8%, FPR: 3.3%, Table 6.3) given the error margins recorded. It is noted that kernel assignment appears to outperform hard assignment (82.8% vs 78.5%) but, as for handgun detection, the performance difference lies within the measured error margin.

Kernel-assignment performance using the RIFT descriptor is shown in Figure A.3 where we see a lesser performance when compared to SIFT. Peak recognition occurs for ($K = 1024$, $\sigma = 0.04$) with a true-positive rate of 78.1% and a false-positive rate of 4.6%. This is significantly lower than for handguns (TPR: 86.9%; FPR: 4.7%, Table 6.3) but we again see that correct tuning of both the number of codewords (K) and smoothing parameter (σ) are required to optimize performance



(a) True-positive performance



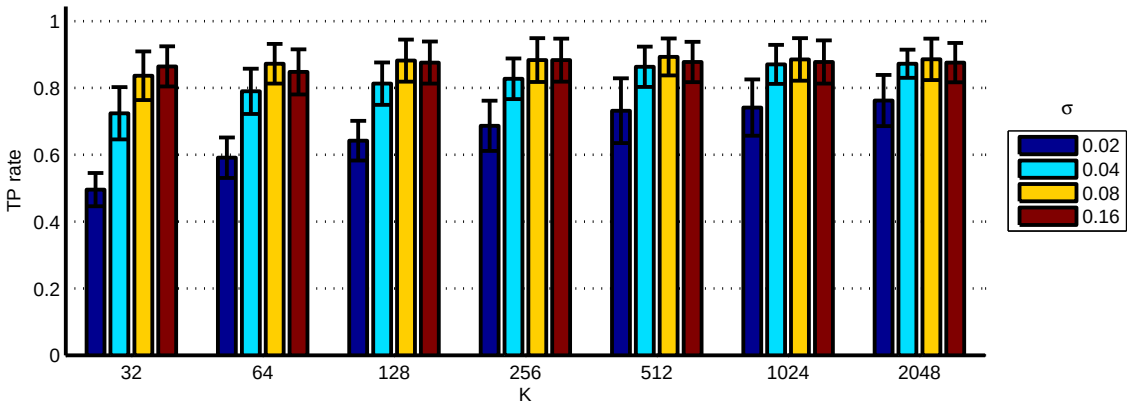
(b) False-positive performance

Figure A.3: Bottle sub-volume results using kernel assignment, SVM classification for RIFT descriptor

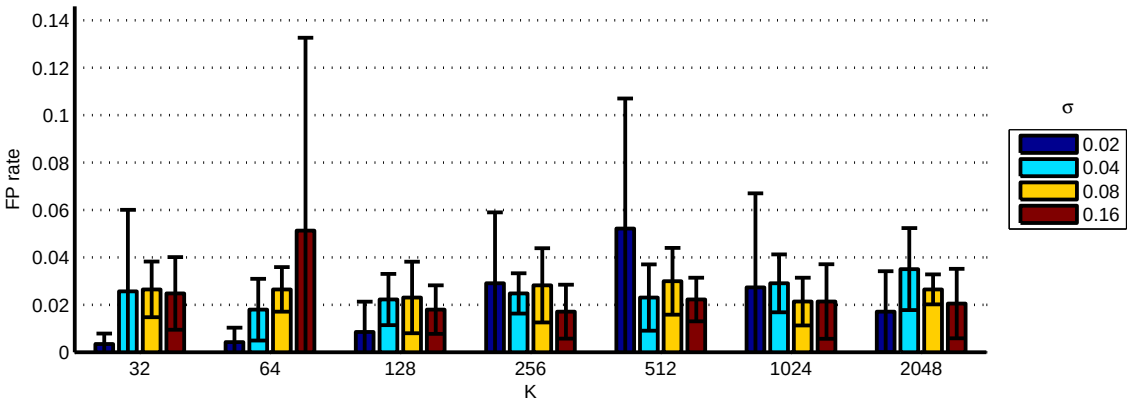
in line with earlier work (van Gemert et al., 2010; Philbin et al., 2008).

Performance for the kernel-assignment method when using the density-histogram descriptor is shown in Figure A.4 where we start to see superior detection when compared to SIFT and RIFT. Correct parameter tuning is obvious from Figure A.4a with improving detection as the number of codewords (K) increases and the smoothing parameter (σ) is adjusted. Highest recognition rates are obtained for ($K = 512$, $\sigma = 0.08$) with a value of 89.3% and corresponding false-positive rate of 3.0% though it can be seen from Figure A.4a that similar results are obtained for $K = \{256, 512, 1024, 2048\}$. These results are again lower than for handguns (TPR: 97.3%; FPR: 1.8%; Table 6.3).

The density-histogram descriptor performance using the kernel-assignment method is shown in Figure A.5 which shows poor performance when the smoothing parameter is too small ($\sigma = 0.02$), in line with the handgun results (Figure 6.15). Peak detection occurs for ($K = 2048$, $\sigma = 0.04$) with a value of 84.4% and corresponding

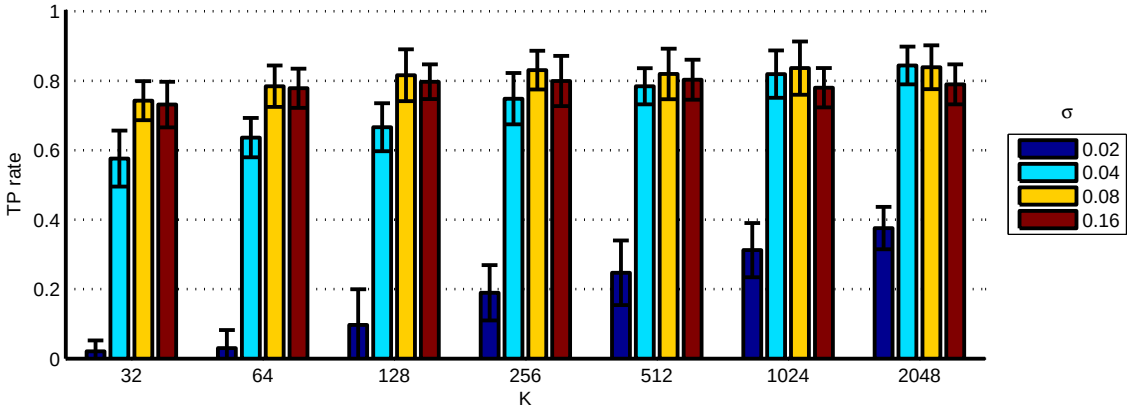


(a) True-positive performance

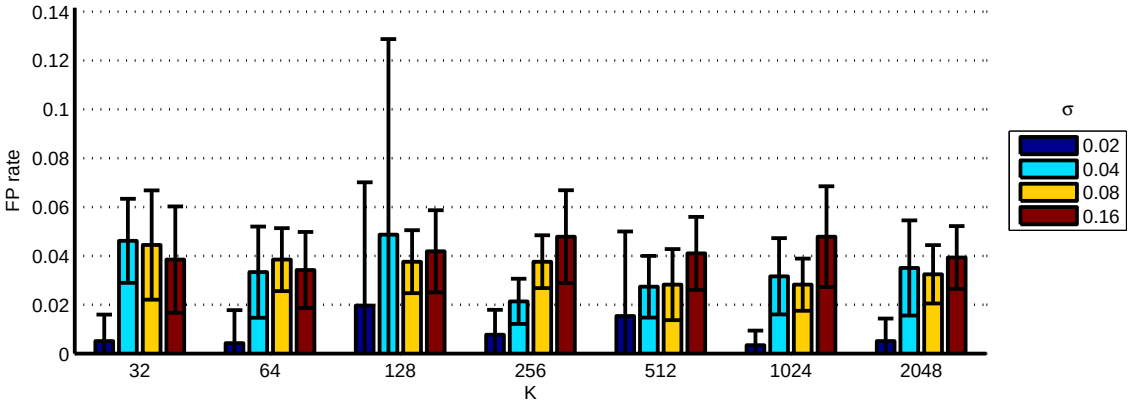


(b) False-positive performance

Figure A.4: Bottle sub-volume results using kernel assignment, SVM classification for density-histogram descriptor



(a) True-positive performance



(b) False-positive performance

Figure A.5: Bottle sub-volume results using kernel assignment, SVM classification for density-gradient-histogram descriptor

false-positive rate of 3.5%.

The best bottle detection results using kernel assignment for each descriptor are summarized in Table A.2. In every case the peak detection rate is lower than for handgun detection (Table 6.3) and the false-positive rate is higher. However, we again see that kernel assignment (when optimally tuned) outperforms hard assignment (Table A.1) in line with existing work (van Gemert et al., 2010; Philbin et al., 2008).

A.3 Uncertainty assignment

For the investigation of uncertainty assignment we use:

$$\sigma = \{0.005,\ 0.01,\ 0.02,\ 0.04,\ 0.08,\ 0.16\}$$

$$K = \{32,\ 64,\ 128,\ 256,\ 512,\ 1024,\ 2048\}$$

Descriptor	K	σ	TP rate (%)	FP rate (%)
SIFT	2048	0.16	82.8 ± 5.7	5.6 ± 1.5
RIFT	1024	0.04	78.1 ± 5.5	4.6 ± 1.6
DH	512	0.08	89.3 ± 5.5	3.0 ± 1.4
DGH	2048	0.04	84.4 ± 5.4	3.5 ± 2.0

Table A.2: Bottle sub-volumes: best detection rate for each descriptor using kernel assignment with SVM classifier

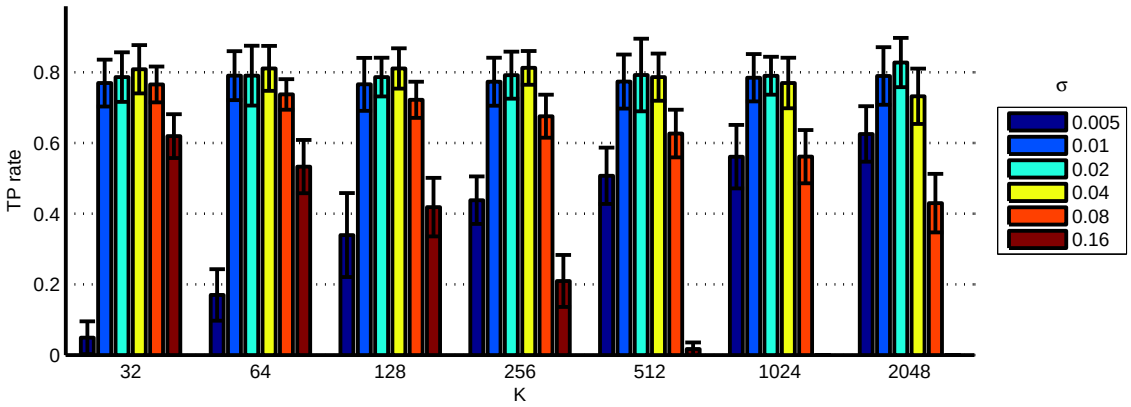
Results for the SIFT descriptor as shown in Figure A.6 where we see a peak detection rate of 82.7% (Figure A.6a) for ($K = 2048, \sigma = 0.02$) with a corresponding false-positive rate of 4.2% (Figure A.6b). These results are very similar to those obtained using kernel assignment (Table A.2) but again show a lesser performance when compared to handgun detection (TPR: 87.0%; FPR: 3.8%; Table 6.4). In Figure A.6a we can see that, for each setting of K, the smoothing parameter (σ) needs to be correctly adjusted: too large (≥ 0.08) or too small (≤ 0.01) and the recognition rate is adversely affected.

The performance of uncertainty assignment using the RIFT descriptor is shown in Figure A.3 with peak recognition of bottles of 78.2% occurring for ($K = 2048, \sigma = 0.01$). The corresponding false-positive rate is 5.6% which again indicates a similar level of performance to kernel assignment (Table A.2). Again we see handgun performance is greater in this configuration (TPR: 87.3%; FPR: 5.1%; Table 6.4).

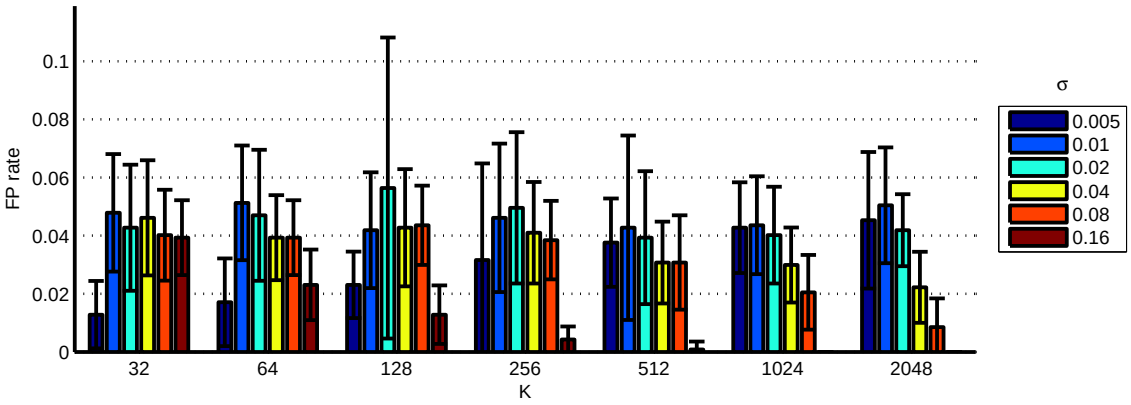
Results for the density histogram are shown in Figure A.8 where we can see more stable results in comparison to the results obtained for SIFT (Figure A.6) and RIFT (Figure A.7) as the codebook size (K) and smoothing parameter (σ) are varied. A higher detection rate is observed with a peak value of 88.2% for ($K = 512, \sigma = 0.04$). Similar results are obtained for all other values of K in the range (86.9%, 87.8%) although the value for σ is adjusted during this process. The false-positive rate is 2.2% for peak recognition.

The density-gradient histogram results are shown in Figure A.9 showing peak detection of 87.2% for ($K = 512$, $\sigma = 0.04$) with a false-positive rate of 4.0%. Selection of values for K and σ need more care than for the density-histogram descriptor as can be seen in Figure A.9a.

Table A.3 shows comparative results for each descriptor with the parametric values used where we can see the DH descriptor yielding the highest true positive and lowest false-positive results. The DGH descriptor has a similar true-positive rate but a noticeably higher false-positive rate. The SIFT and RIFT descriptors lag behind as before.



(a) True-positive performance

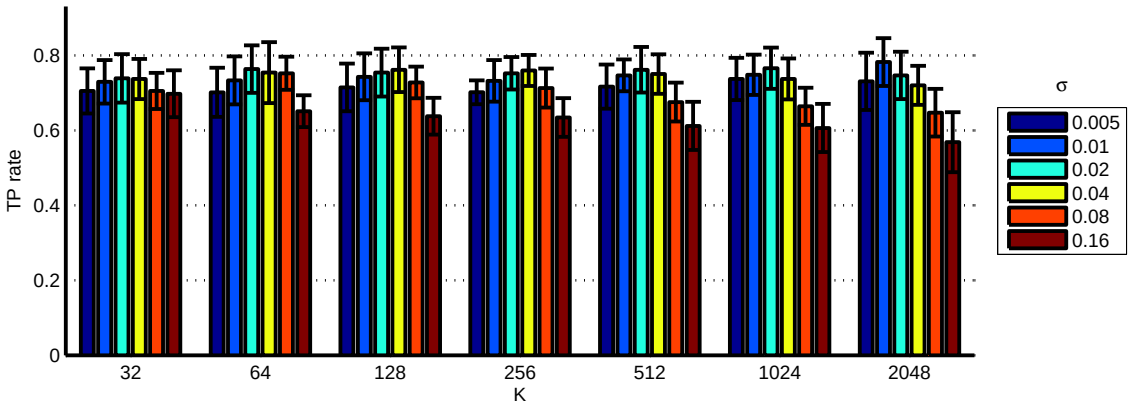


(b) False-positive performance

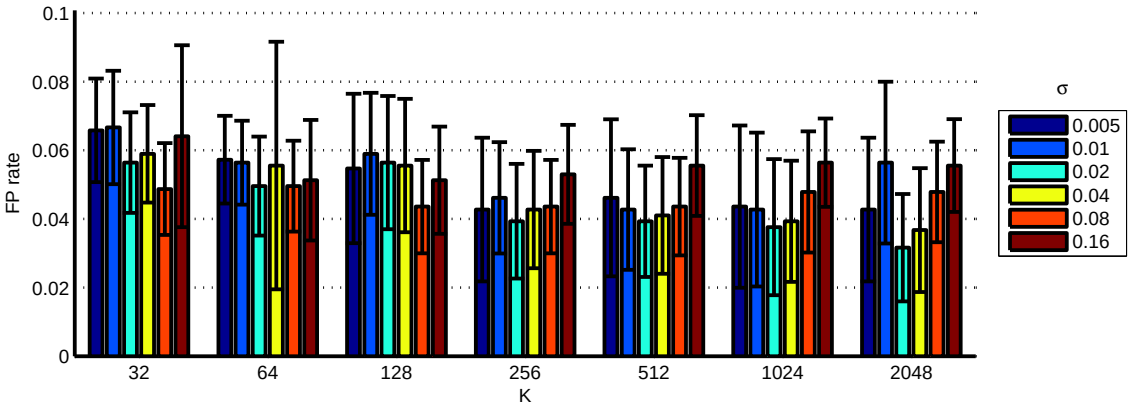
Figure A.6: Bottle sub-volume results using uncertainty assignment, SVM classification for SIFT descriptor

Descriptor	K	σ	TP rate (%)	FP rate (%)
SIFT	2048	0.02	82.7 ± 7.0	4.2 ± 1.2
RIFT	2048	0.01	78.2 ± 6.4	5.6 ± 2.4
DH	512	0.04	88.2 ± 4.7	2.2 ± 1.3
DGH	512	0.04	87.2 ± 6.8	4.0 ± 1.8

Table A.3: Optimized bottle detection rate for each descriptor using uncertainty assignment with SVM classifier

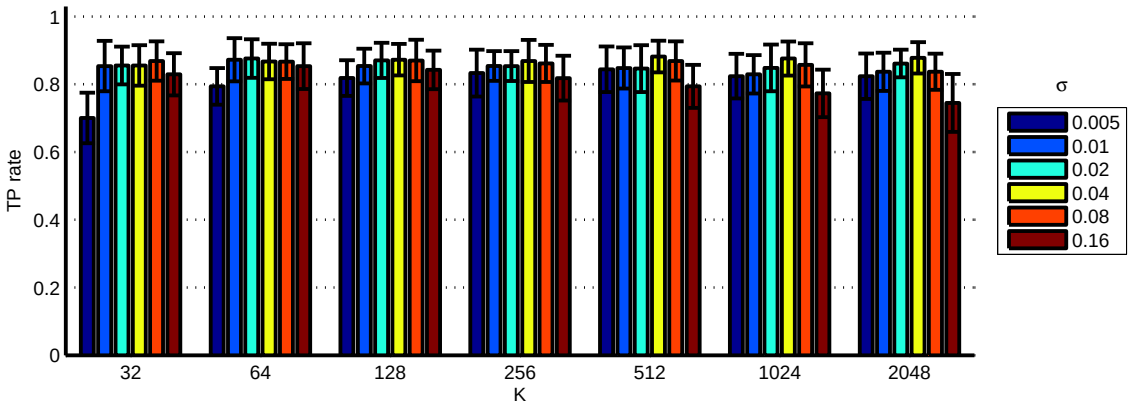


(a) True-positive performance

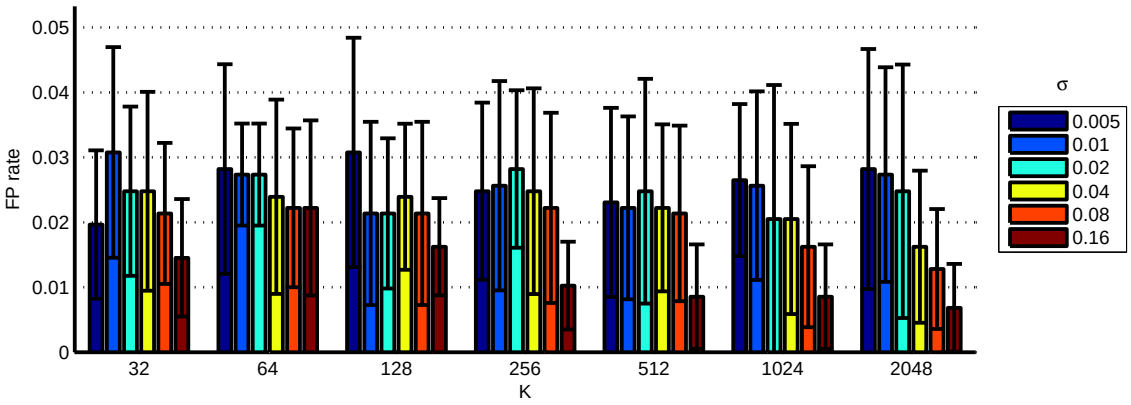


(b) False-positive performance

Figure A.7: Bottle sub-volume results using uncertainty assignment, SVM classification for RIFT descriptor

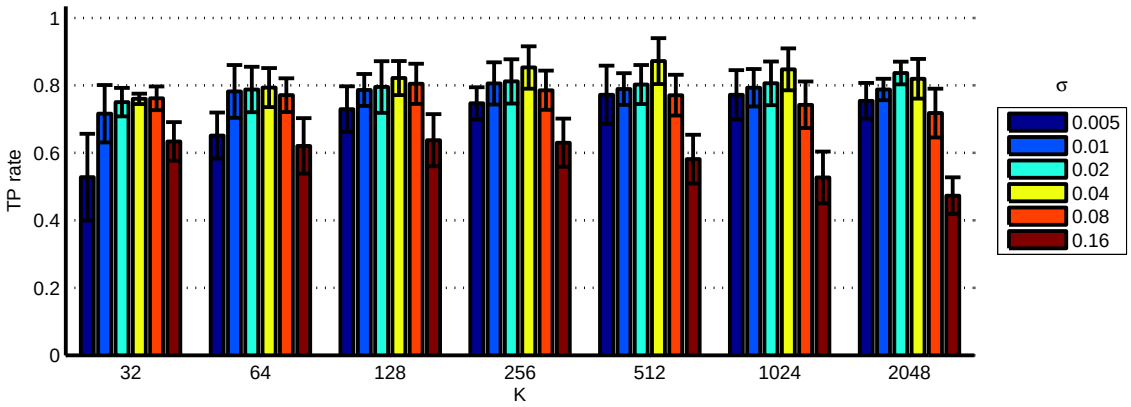


(a) True-positive performance

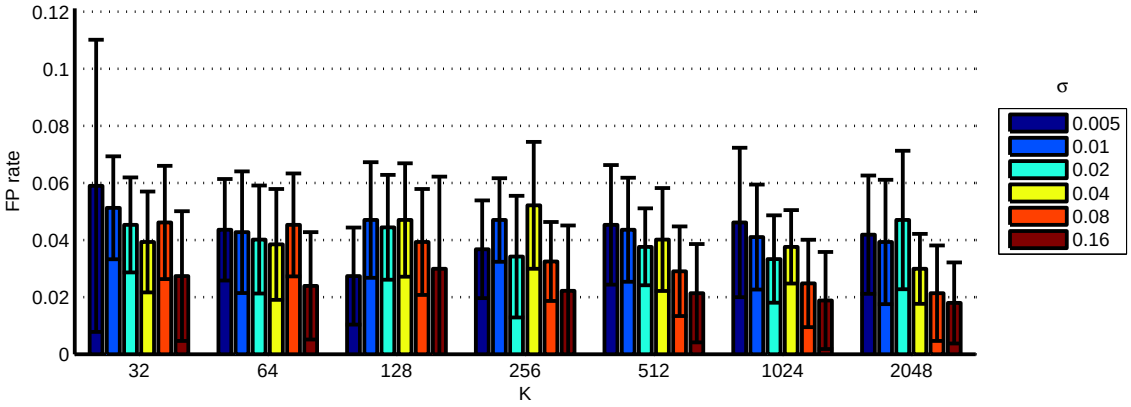


(b) False-positive performance

Figure A.8: Bottle sub-volume results using uncertainty assignment, SVM classification for density-histogram descriptor



(a) True-positive performance



(b) False-positive performance

Figure A.9: Bottle sub-volume results using uncertainty assignment, SVM classification for density-gradient-histogram descriptor

Appendix B

Codebook: handgun whole-bag results

Rather than examine detection using cropped threat sub-volumes and clutter we can now evaluate performance using whole-baggage volumes as the source data - a task that is more akin to that undertaken by security personnel at airports (Shanks and Bradley, 2004). It is anticipated that this will be a harder undertaking than before as the amount of clutter in the threat volumes is greatly increased.

A small dataset is used for this work and details are given in Table B.1.

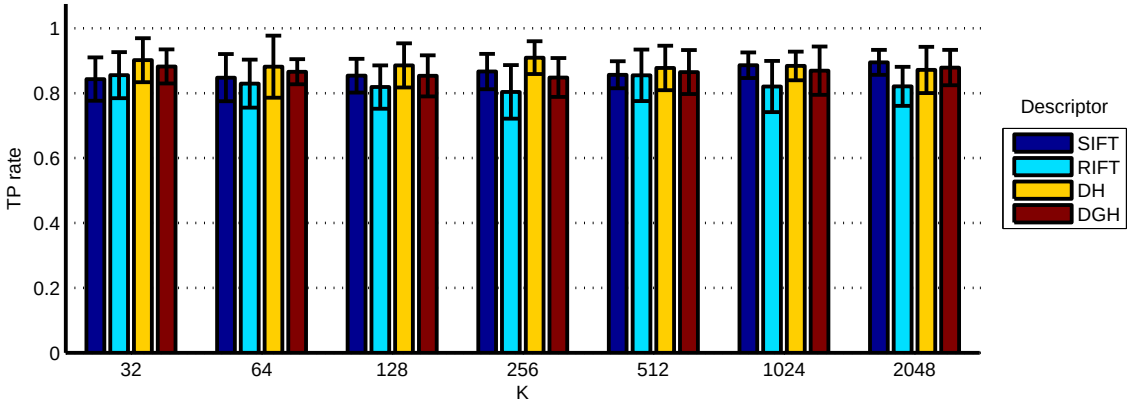
B.1 Hard assignment

Hard-assignment results are shown in Figure B.1. Figure B.1a shows the detection rates for each descriptor as we vary the number of codewords (K). We can see that the descriptors have similar detection rates with density histogram having the highest rate of 90.9%, closely followed by SIFT (89.5%) and density-gradient histogram (88.2%) with RIFT at 85.6%. There is a clearer distinction in the false-positive performance with density-gradient histogram having the best performance at 16.1% for 1024 codewords. The optimum detection rates are given in Table B.2.

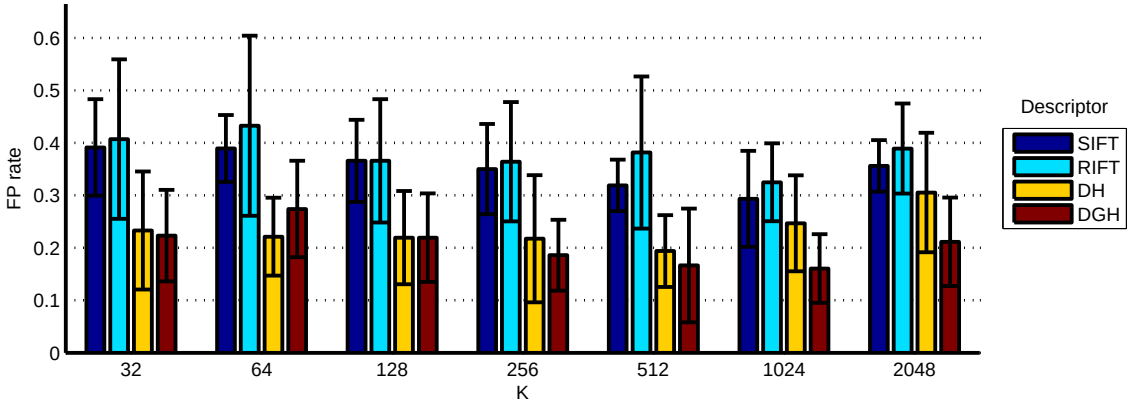
The degree of uncertainty in these results prevents a clear result on the true-positive results but the false-positive results seem to show an advantage of density histogram and density-gradient histogram over SIFT and RIFT.

Item	Quantity
Threat	306
Clear	179

Table B.1: Handgun Whole-Baggage Dataset



(a) True-positive performance

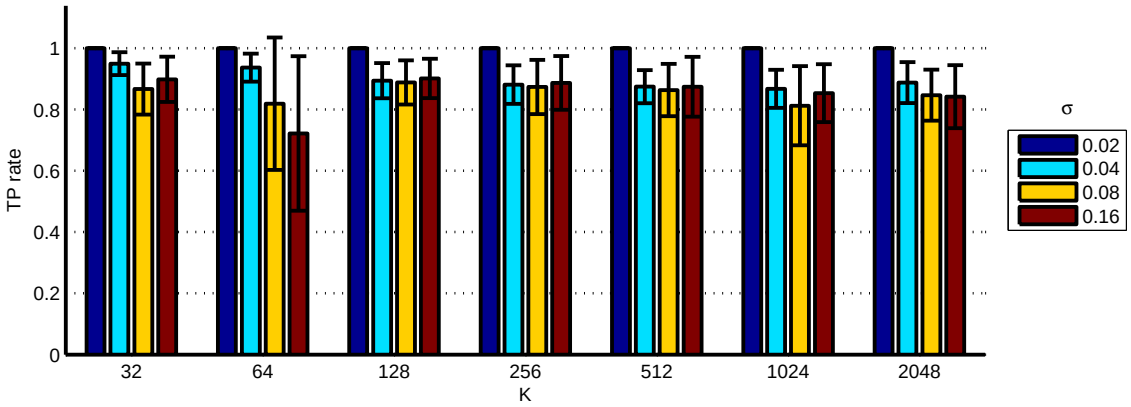


(b) False-positive performance

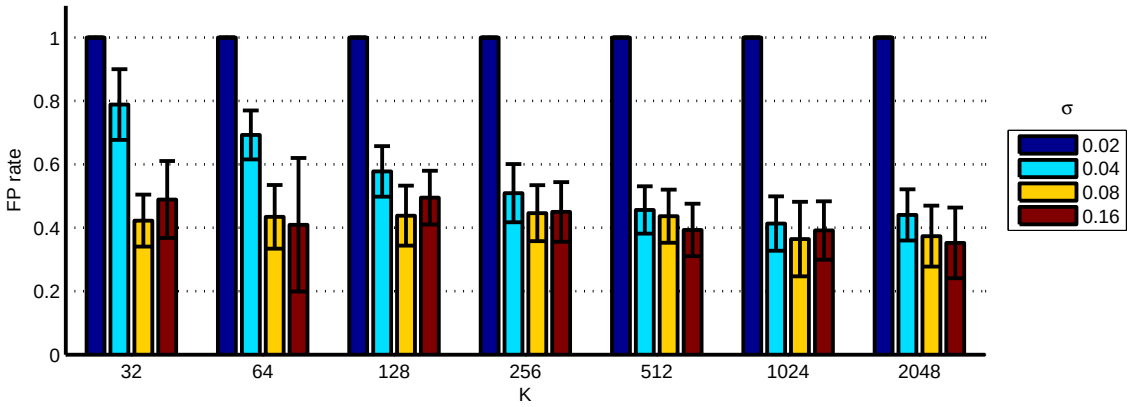
Figure B.1: Whole-volume handgun results using hard assignment and SVM classification

Descriptor	K	TP rate (%)	FP rate (%)
SIFT	2048	89.5 ± 3.9	35.6 ± 4.9
RIFT	32	85.6 ± 7.1	40.7 ± 15.2
DH	256	90.9 ± 5.0	21.7 ± 12.1
DGH	32	88.2 ± 5.2	22.3 ± 8.7

Table B.2: Whole-volume handgun detection rates for each descriptor using SVM



(a) True-positive performance

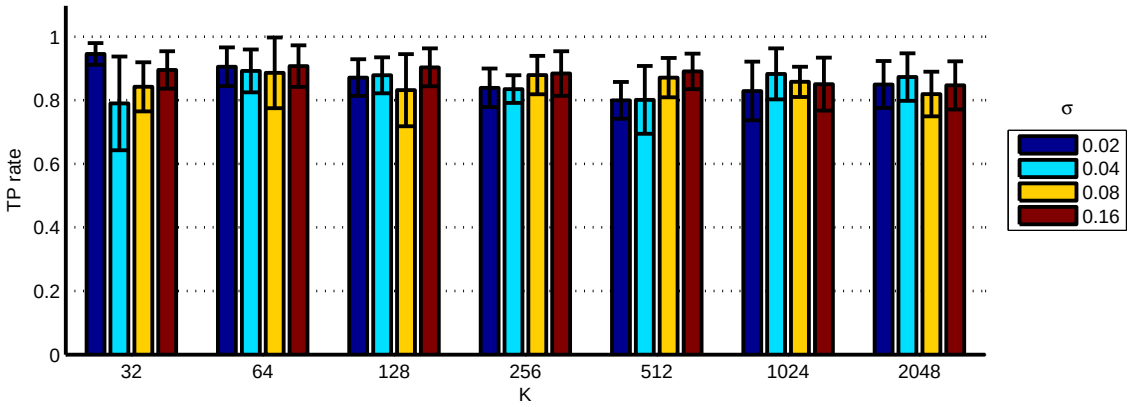


(b) False-positive performance

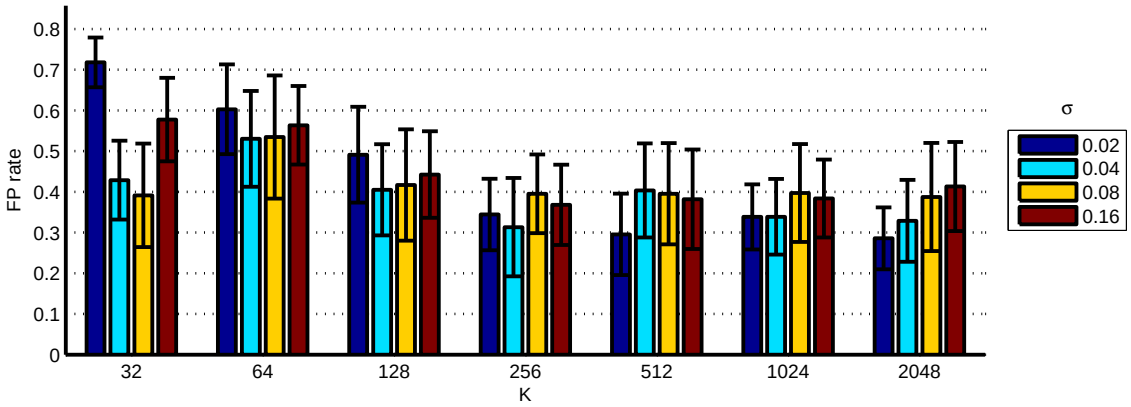
Figure B.2: Whole-volume handgun results using kernel assignment and SVM classification for SIFT descriptor

B.2 Kernel assignment

Kernel-assignment results using the SIFT descriptor are shown in Figure B.2. For $\sigma = 0.02$ both true-positive and false-positive rates are 100% indicating that the SVM has been completely failed to distinguish between threat and non-threat items. For $\sigma = 0.04$ we can see that the the true-positive and false-positive results are tending to 100% as the number of clusters is reduced. For example when $K = 32$ we have a detection rate of 94.9% but a high false-positive rate of 78.9%. When considering both the true-positive performance (Figure B.2a) and the false-positive performance (Figure B.2b) we can again see that best performance is achieved with a setting of $\sigma = \{0.08, 0.16\}$. The best detection rate under these conditions is 90.1% and occurs for ($K = 128$, $\sigma = 0.16$) with a corresponding false-positive rate of 49.5%. The lowest false-positive rate is 35.2% and occurs with a setting of ($K = 2048$, $\sigma = 0.16$) for which the true-positive rate is 84.1%.



(a) True-positive performance

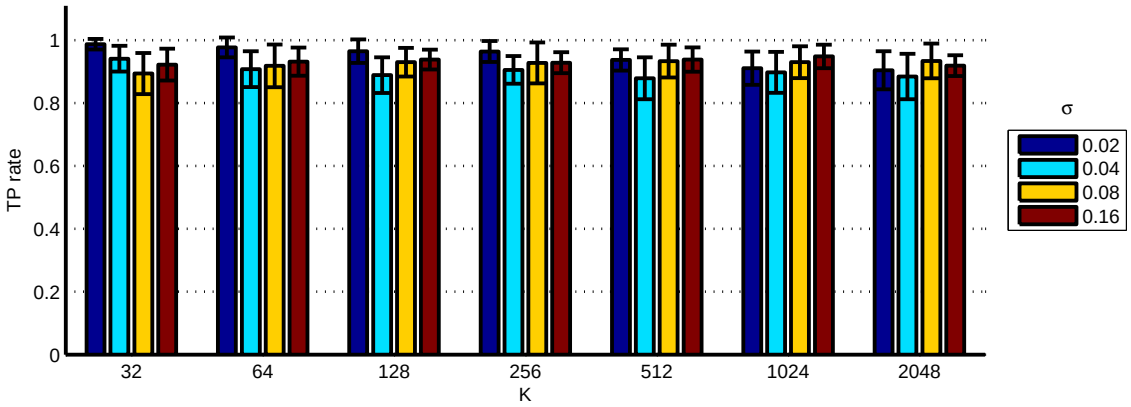


(b) False-positive performance

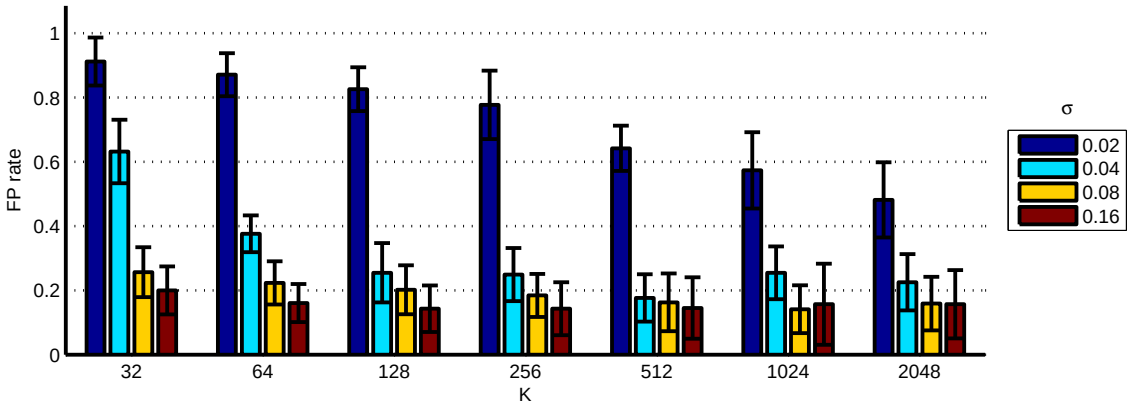
Figure B.3: Whole-volume handgun results using kernel assignment and SVM classification for RIFT descriptor

Using the RIFT descriptor with kernel-assignment produces the results in Figure B.3. We can see that the number of visual words (K) noticeably affects the false-positive rate with an increase as K is reduced. As similar effect is seen to SIFT where the best overall performance is achieved for $\sigma = \{0.08, 0.16\}$. Under these settings the highest detection occurs when ($K = 64$, $\sigma = 0.16$) with a value of 90.7% and corresponding false-positive rate of 56.4%. If we choose to optimize the false-positive rate then the best performance occurs for ($K = 512$, $\sigma = 0.02$) with a detection rate of 79.9% and false-positive rate of 29.5%.

Density-histogram results using kernel assignment are shown in Figure B.4 where we see a now familiar pattern in the false-positive results when the smoothing parameter (σ) is too small. In this case $\sigma = 0.02$ produces a false-positive rate that is noticeably high for all value of K and using $\sigma = 0.04$ produces high results for $K < 128$. If we ignore these settings then the best detection rate occurs for ($K = 1024$, $\sigma = 0.16$) with a rate of 94.8% and corresponding false-positive rate of



(a) True-positive performance

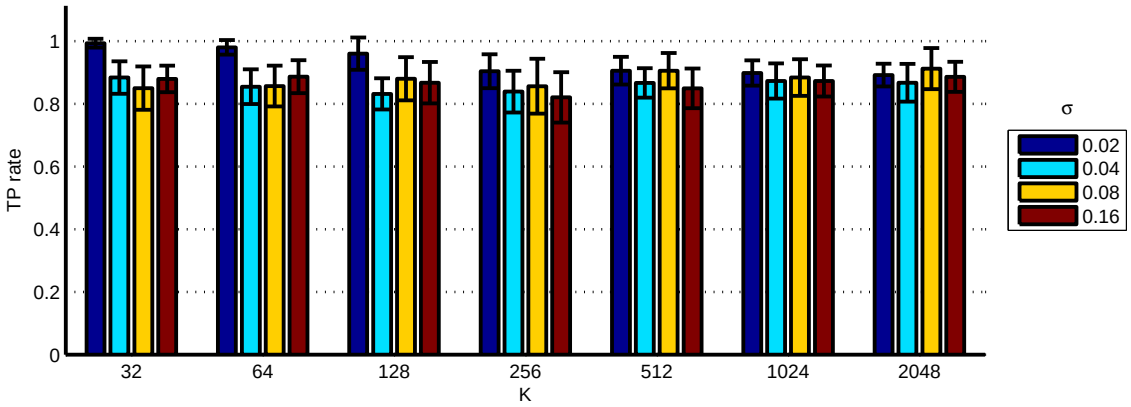


(b) False-positive performance

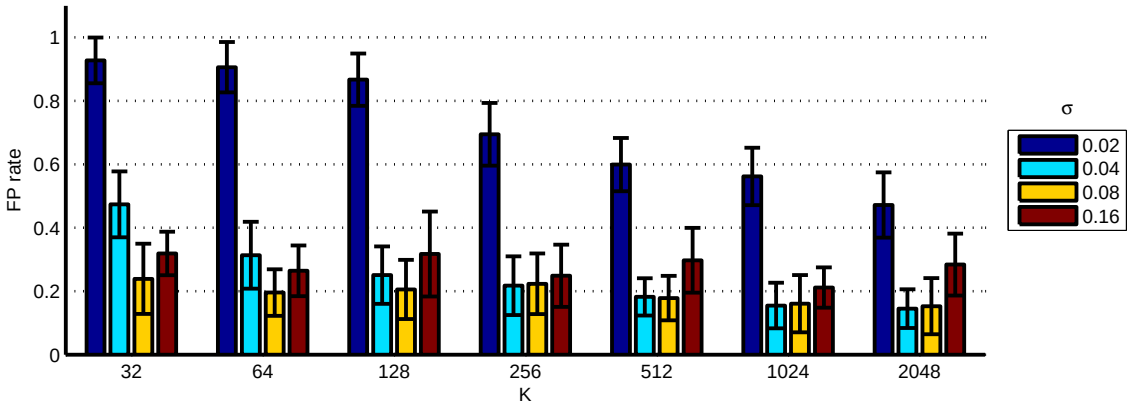
Figure B.4: Whole-volume handgun results using kernel assignment and SVM classification for density-histogram descriptor

15.7%. The best false-positive rate is achieved using the same number of clusters ($K = 1024$) but setting the smoothing parameter to $\sigma = 0.08$ for which we see a detection rate of 93.0% and false-positive rate of 14.1%. Given the measurement error, this reduced false-positive rate does not represent a clear improvement over that which was achieved for the highest detection rate.

Density-gradient histogram results using kernel-assignment are shown in Figure B.5 where we again see the same pattern in the false-positive results when the smoothing parameter (σ) is too small. As for density histogram, we ignore all results when $\sigma = 0.02$ and for $\sigma = 0.04$ when $K < 128$. Under these conditions the best detection rate occurs for ($K = 2048$, $\sigma = 0.08$) with a rate of 91.2% and corresponding false-positive rate of 15.3%. Best false-positive rate is achieved using ($K = 2048$, $\sigma = 0.04$) where we see a detection rate of 86.7% and false-positive rate of 14.5%. Again the measurement error does not allow us to declare an optimum setting of smoothing parameter or number of clusters to achieve best detection or



(a) True-positive performance



(b) False-positive performance

Figure B.5: Whole-volume handgun results using kernel assignment and SVM classification for density-gradient-histogram descriptor

minimum false detection.

The best detection results for each descriptor are summarized in Table B.3. From this we can make a number of observations. Firstly, all descriptors yield detection rates above 90.0% with density histogram at the top with 94.8%. However, the measurement uncertainty in each case allows all the measurements to overlap. SIFT and RIFT lag behind the other descriptors based on the mean results for both true positive and false positive measures. There is a noticeable difference in false-positive performance with SIFT and RIFT significantly higher than DH and DGH. We also note that the best SIFT and RIFT performance is achieved with a relatively small number of visual words ($K = \{64, 128\}$) whereas the performance for DH and DGH uses many more ($K = \{1024, 2048\}$). In all cases the smoothing parameter, σ, is at the higher end of the tested parameter values ($\sigma = \{0.08, 0.16\}$).

Descriptor	K	σ	TP rate (%)	FP rate (%)
SIFT	128	0.16	90.1 ± 6.4	49.5 ± 8.5
RIFT	64	0.16	90.7 ± 6.5	56.4 ± 9.7
DH	1024	0.16	94.8 ± 3.7	15.7 ± 12.6
DGH	2048	0.08	91.2 ± 6.6	15.3 ± 8.8

Table B.3: Best whole-volume handgun detection rate for each descriptor using kernel assignment with SVM classifier

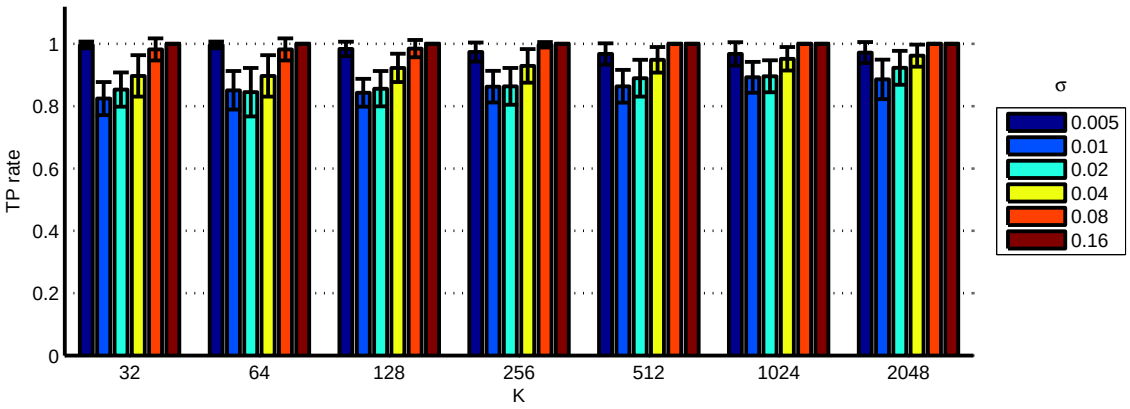
B.3 Uncertainty assignment

Uncertainty-assignment results using the SIFT descriptor are shown in Figure B.2. It can be see that tuning of the smoothing parameter is required as poor results are achieved if it is too low ($\sigma = 0.005$) or too high ($\sigma > 0.04$). It can also be seen that the false-detection performance for ($\sigma = 0.04$) deteriorates as the number of visual words (K) increases. Excluding these settings the highest detection rate is achieved for ($K = 2048$, $\sigma = 0.02$) with a rate of 92.3% and corresponding false-positive rate of 44.8%. The settings for minimum false-positive rate are ($K = 1024$, $\sigma = 0.01$) with a rate of 31.1% and detection rate of 89.2%.

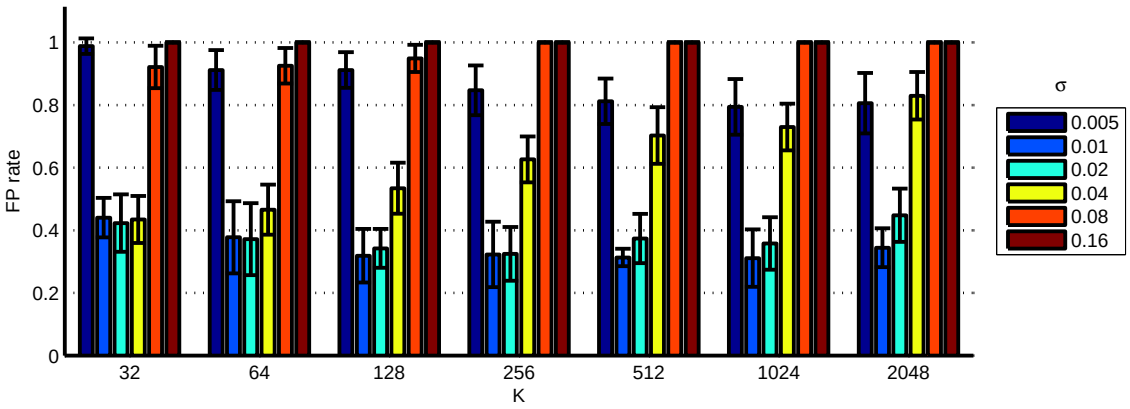
Using the RIFT descriptor with uncertainty assignment produces the results in Figure B.7. A different pattern is seen in this case where the false-positive rate is abnormally high for $\sigma = \{0.08, 0.16\}$ for all values of K and also for ($\sigma = 0.04$, $K > 128$), ($\sigma = 0.02$, $K > 512$) and ($\sigma = 0.01$, $K = 2048$). Excluding these settings from the results we obtain a highest detection rate when ($K = 128$, $\sigma = 0.04$) with a value of 90.2% and corresponding false-positive rate of 46.4%. If we choose to optimize the false-positive rate then the best performance occurs for ($K = 128$, $\sigma = 0.01$) with a detection rate of 80.8% and false-positive rate of 33.3%.

Density-histogram results using uncertainty assignment are shown in Figure B.8 where we see an increasingly poor overall performance as the number of visual words is increased. If we exclude regions where the false-positive rate exceeds 40.0% we find that the best detection occurs for a setting of ($K = 1024$, $\sigma = 0.04$) with a rate of 92.2% and corresponding false-positive rate of 30.2%. Tuning for lowest false-positive rate yields a setting of ($K = 256$, $\sigma = 0.02$) with a true-positive rate of 91.0% and false-positive rate of 14.9%.

Density-gradient-histogram results using uncertainty assignment are shown in Figure B.9 where we see false-positive rate increase as a combination of the number of visual words and smoothing parameter setting. Using a smoothing parameter setting of $\sigma = 0.16$ produces poor false positive results for all values of K. Excluding settings that produce a false-positive rate above 40.0% we see the best detection

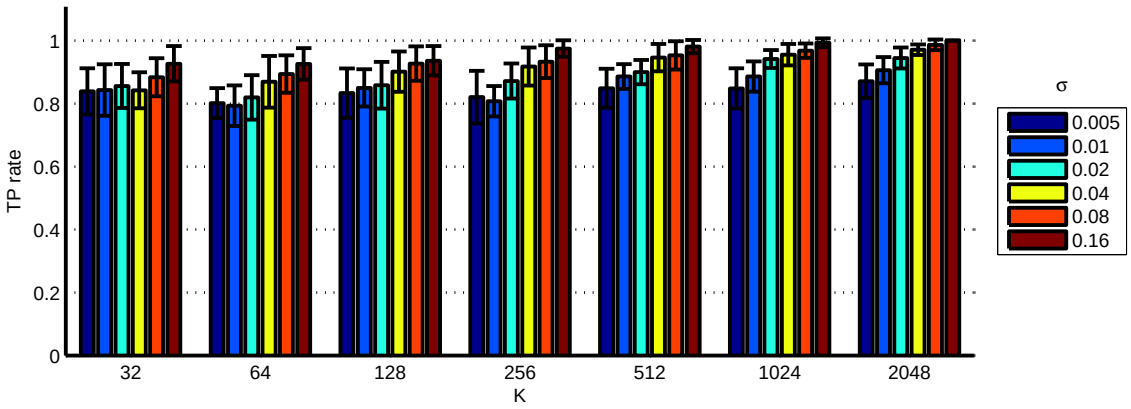


(a) True-positive performance

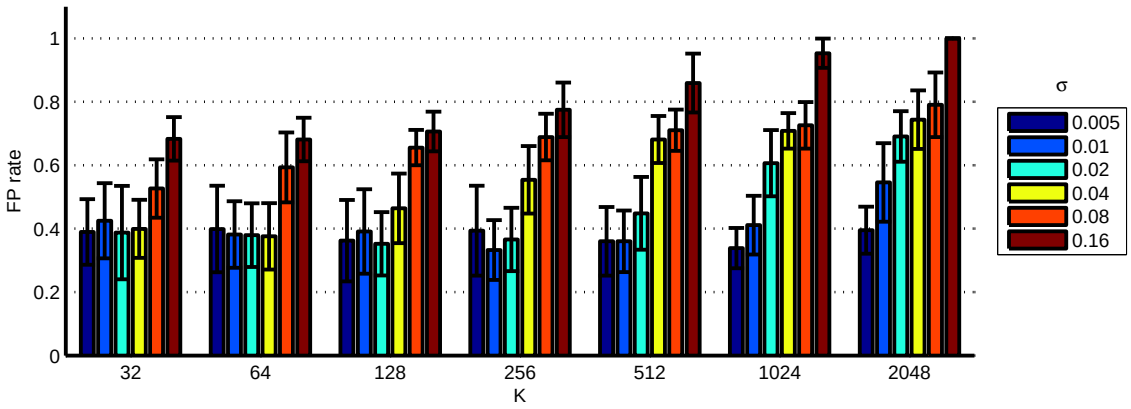


(b) False-positive performance

Figure B.6: Whole-volume handgun results using kernel assignment and SVM classification for SIFT descriptor

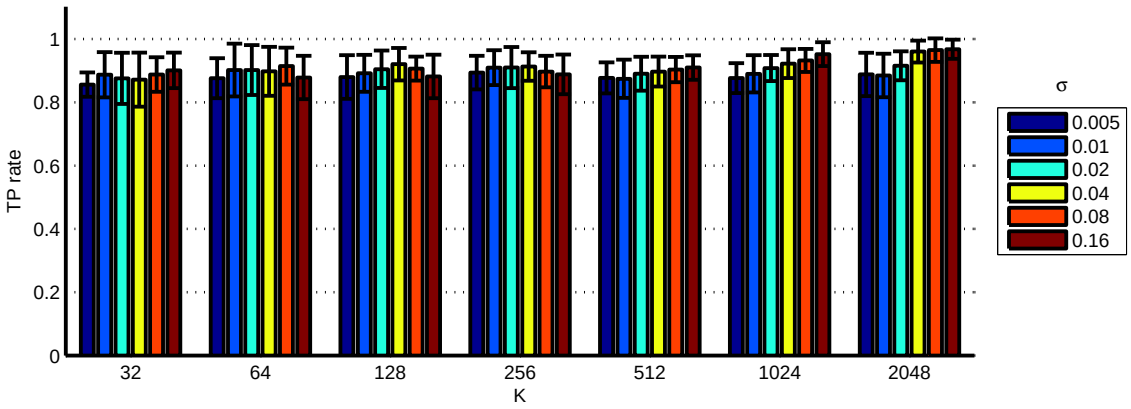


(a) True-positive performance

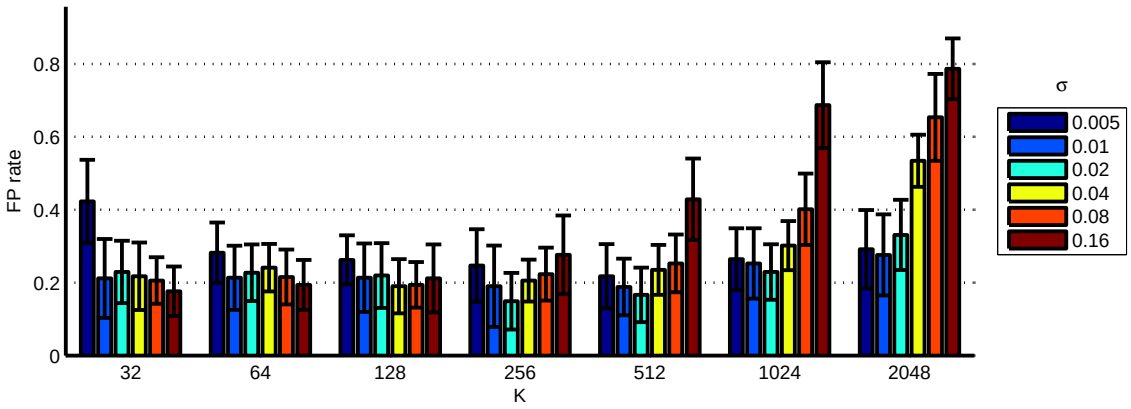


(b) False-positive performance

Figure B.7: Whole-volume handgun results using uncertainty assignment and SVM classification for RIFT descriptor

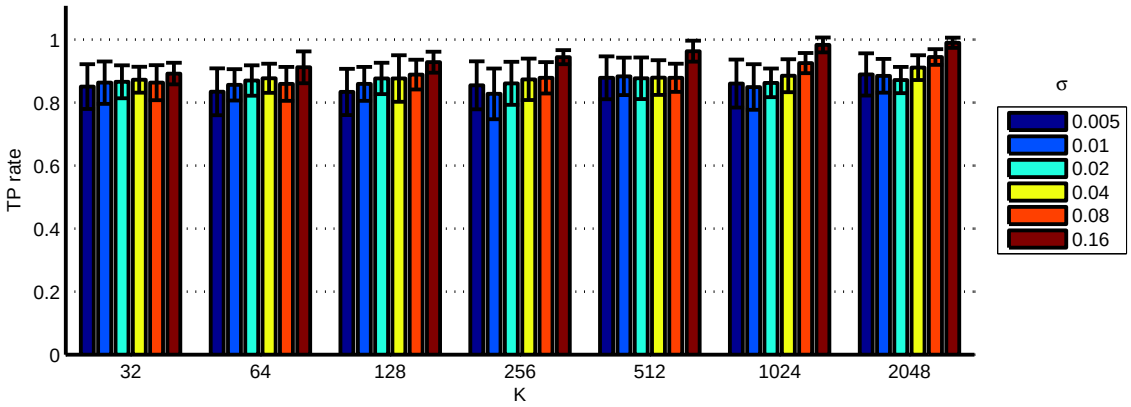


(a) True-positive performance

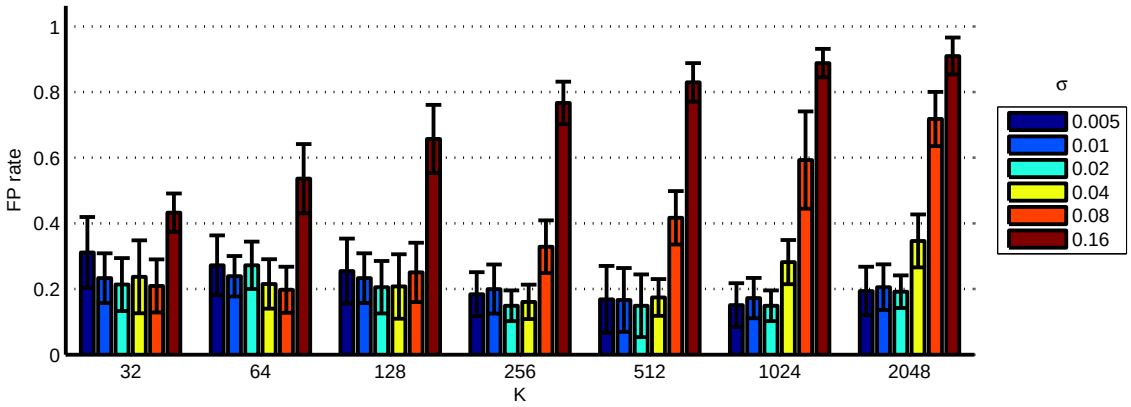


(b) False-positive performance

Figure B.8: Whole-volume handgun results using uncertainty assignment and SVM classification for density-histogram descriptor



(a) True-positive performance



(b) False-positive performance

Figure B.9: Whole-volume handgun results using uncertainty assignment and SVM classification for density-gradient histogram descriptor

occurring for a setting of ($K = 2048$, $\sigma = 0.04$) when the detection rate is 91.1% and the false-positive rate is 34.6%. If we choose the setting for lowest false-positive rate we get a setting of ($K = 1024$, $\sigma = 0.02$) with a detection rate of 86.3% and false-positive rate of 14.9%.

The best detection results for each descriptor are summarized in Table B.4. From this we can make a number of observations. Firstly, all descriptors yield detection rates above 90.0% with SIFT at the top with 92.3%. However, the measurement uncertainty in each case allows all the measurements to overlap. Again there is a noticeable difference in false-positive performance with SIFT and RIFT significantly higher than DH and DGH. This time the smoothing parameter, σ, is in the middle of the tested settings ($\sigma = \{0.02, 0.04\}$) suggesting that the assignment normalization achieved using the uncertainty approach is optimal for settings at or below the mean adjacent cluster distance (Section 6.5.1).

Descriptor	K	σ	TP rate (%)	FP rate (%)
SIFT	2048	0.02	92.3 ± 5.4	44.8 ± 8.5
RIFT	128	0.04	90.2 ± 6.4	46.4 ± 11.0
DH	1024	0.04	92.2 ± 4.5	30.2 ± 6.7
DGH	2048	0.04	91.1 ± 4.0	34.6 ± 8.1

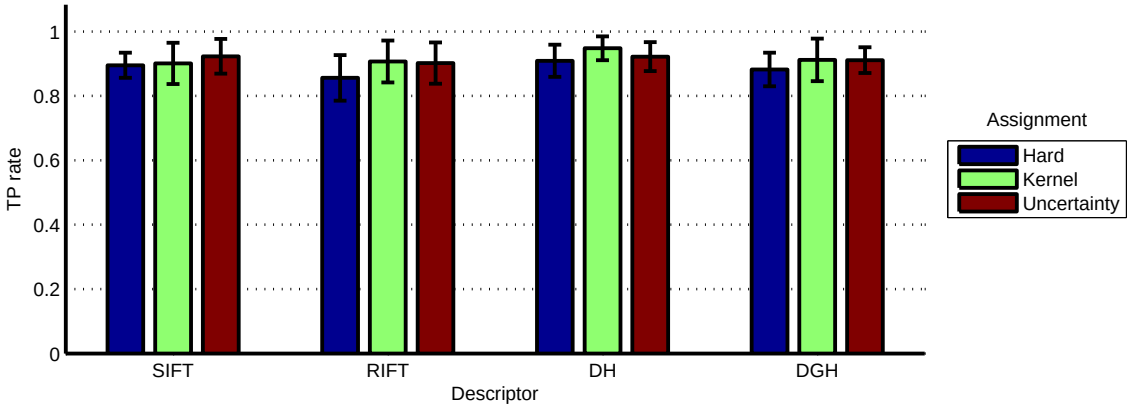
Table B.4: Best whole-volume handgun detection rate for each descriptor using uncertainty assignment with SVM classifier

Descriptor	Assignment method	K	σ	TP rate (%)	FP rate (%)
SIFT	Uncertainty	2048	0.02	92.3 ± 5.4	44.8 ± 8.5
RIFT	Kernel	64	0.16	90.7 ± 6.5	56.4 ± 9.7
DH	Kernel	1024	0.16	94.8 ± 3.7	15.7 ± 12.6
DGH	Kernel	2048	0.08	91.2 ± 6.6	15.3 ± 8.8

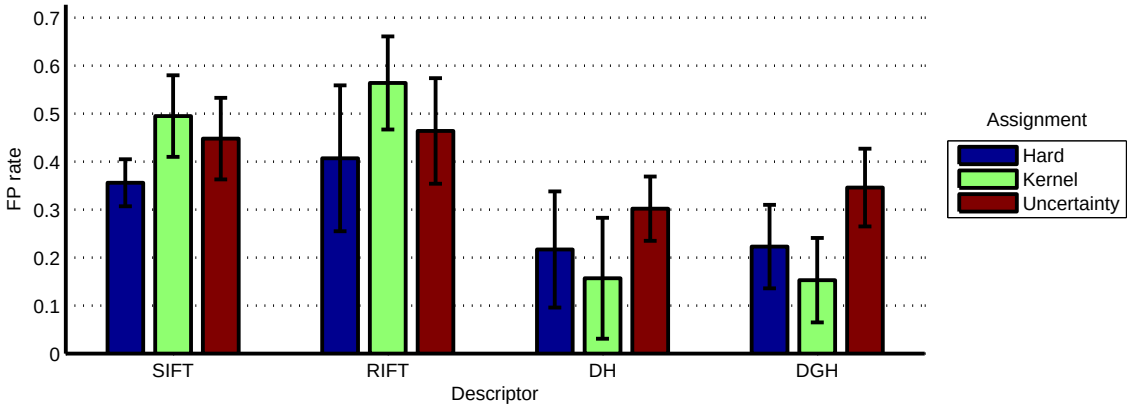
Table B.5: Whole-volume handgun best detection results and settings

B.4 Summary of performance

We can now summarize and compare the performance of the four descriptors with the three assignment methods when analyzing whole baggage items. Figure B.10 shows how each descriptor performs for each assignment method. Figure B.10a shows the detection performance where we can see the general outperformance of density histogram over density-gradient histogram, SIFT and RIFT. The best detection is obtained using kernel-assignment with the density-histogram descriptor: 94.8% true-positive rate. This setting also yields a relatively low false-positive result: 15.7%. SIFT and RIFT have significantly higher false-positive rates when compared to DH and DGH. Table B.5 summarizes the best performing result for each descriptor where we can see that the density-histogram descriptor has the highest overall detection result (94.8%) with almost the lowest false-positive rate (15.7%).



(a) True-positive performance



(b) False-positive performance

Figure B.10: Best detection whole-volume handgun results summary using SVM classification

Appendix C

Classification measures

In the analysis of results within this thesis we make use of a number of methods. As an example, consider the recognition of a pistol within a set of baggage items. For each baggage item a decision is made as to whether a pistol is present (positive) or not (negative). Some of the positive decisions will be correct (true positive, *tp*) and some incorrect (false positive, *fp*). Similarly some of the negative decisions will be correct (true negative, *tn*) and some incorrect (false negative, *fn*).

Consider the recognition of bags containing pistols outlined in Table C.1. Table C.1a shows the number of clutter bags and pistol bags together with the recognition totals for each type. Table C.1b transforms this data into true/false positive/negative terms.

From this example we can see the following:

$tp + fn$ =the total number of bags containing the target item

$tn + fp$ =the total number of clutter bags

Various statistical results can be calculated from this data, as shown in Table C.2. The true-positive rate measures the fraction of bags containing pistols that were correctly identified. The false-positive rate measures the fraction of clutter bags that were incorrectly marked as containing a pistol. Recall has the same formulation as true-positive rate, whereas precision records the fraction of bags declared as containing a pistol that *actually* do contain a pistol.

We can see from this that, for example, the false-negative rate (FNR) is closely related to the true-positive rate (TPR):

$$FNR = 1 - TPR$$

	Total	Correctly identified	Incorrectly identified
Clutter Bags	456	453	3
Pistol Bags	43	41	2

(a) Baggage data

	Positive	Negative
True	41	453
False	3	2

(b) Rewritten in true/false, positive/negative terms

Table C.1: Example recognition performance

Name	Formulation
True-positive rate, TPR	$\frac{tp}{tp+fn}$
False-positive rate, FPR	$\frac{fp}{tn+fp}$
True negative rate, TNR	$\frac{tn}{tn+fp}$
False-negative rate, FNR	$\frac{fn}{tp+fn}$
Precision	$\frac{tp}{tp+fp}$
Recall	$\frac{tp}{tp+fn}$

Table C.2: Classification measures